Architectures for Data Standardization and Interoperability in Patient Centered Outcomes Research Institute Clinical Research Data Networks

James R. Campbell, MD[1], Lemuel R. Waitman, PhD[2], Abel N. Kho MD MS[3], Thomas R. Campion, Jr. PhD[4], Samuel T. Rosenbloom, MD, MPH[5]

[1]University of Nebraska Medical Center, Omaha, NE, [2]University of Kansas Medical Center, Kansas City, KS, [3]Northwestern University Feinberg School of Medicine, Chicago, IL, [4]Weill Cornell Medical College, New York, NY, [5]Vanderbilt University Medical Center, Nashville, TN

Abstract

The Patient Centered Outcomes Research Institute's (PCORI) Clinical Data Research Network (CDRN) initiative promises to test the reusability of electronic health records (EHR) and other data sources to support comparative effectiveness research. This effort is concurrent with a national investment in EHRs compliant with Nationwide Health Information Network (NwHIN) standards designed to develop interoperable shared data. The interoperation and utility of this data is untested by clinical research at a national scale. This panel will bring together four recently funded CDRNs who will describe the approaches to interoperability, data models, and standardization they are incorporating in their network.

Description

The vision of the Nationwide Health Information Network incorporates an expectation that data collected clinically in the Electronic Health Record will be freely shared for Public Health and Patient Care Research. Although the Office of the National Coordinator is promoting standards, services and policies to achieve NwHIN goals, many issues remain. A grand challenge to the effective development of collaborative patient-centered outcomes research is the development of a common data model employing NwHIN standards and policies that will effectively support aggregation and analysis of EHR, research and outcomes data sets across and between research institutions.

In April, the Patient Centered Outcomes Research Institute announced plans to build a National Patient Centered Clinical Research Network composed of 1) up to eight Clinical Data Research Networks (CDRN) centered around healthsystems that encompass at least one million people and can follow patient longitudinally and support data-access research; and 2) twelve to eighteen Patient Powered Research Networks (PPRN) composed of patients organized around a condition who are motivated to participate in outcomes research. These CDRN and PPRNs are tasked to create interoperable databases and partner to create the greater national network.

Applicants were charged with addressing 13 review criteria with the second focused on standardization and interoperability. Per the PCORI CDRN review criteria:

Describe current informatics standards, interoperability between systems, and plans for achieving data standardization and interoperability between systems within network and across networks.

- *Describe the clinical and information technology systems at each of the healthcare systems participating in the network, including specific standards used for capture and storage of various clinical data elements (diagnoses, prescriptions, laboratory tests and results, radiologic images, progress notes). Also address issues of adherence to standards within systems.*
- *Describe the standards in use for information exchange between systems. Identify interoperability gaps within and between systems. Present plans for enhancing standardization and interoperability of data within and across the network's component systems during the award period.*
- *Describe the policies, procedures, tools, and methods already in place to ensure that data collected across these systems are comparable and valid for research purposes. Describe plans for further developing such policies, procedures, tools, and techniques during the 18-month award period.*
- *Demonstrate an understanding of the intent of this project to work toward data standardization and interoperability across CDRNS and PPRNs and express willingness to work toward these ends.*

This panel will include presentations by four leaders in informatics who are developers of a PCORI collaborative network. They will present, explain and share their approaches to inter-institutional research data management with attention to issues of:

- Compliance with data standardization and procedures for translation of semantics between disparate sources including EHRs and legacy data sets
- Adherence to federal, state and institutional policies for data management security and confidentiality
- Tooling for data communication, aggregation and analysis across diverse research centers

Discussion will emphasize experience with the utility of their strategies, challenges they have faced, and thorny issues that remain unsolved to create a viable national network.

The moderator will briefly introduce applicable standards proposals, regulations and policies. Each presenter will discuss their data model and tooling with emphasis on their depth of experience and the progress of their network. This will be followed by a discussion with audience participation examining the utility and outstanding issues of the data models discussed.

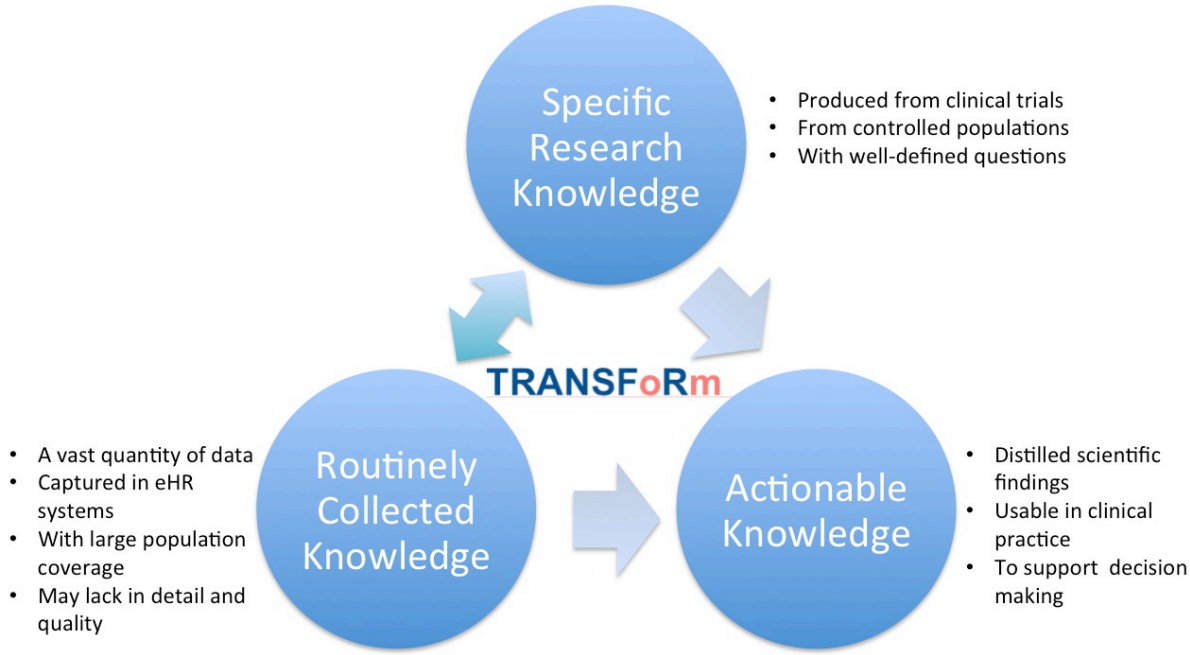
TRANSFoRm digital infrastructure: The architecture for The Learning Healthcare System in Europe

Vasa Curcin PhD[a], Theodoros N Arvanitis[b] DPhil, Piotr Brodka PhD[c], Derek Corrigan MSc[d], Brendan C Delaney MD[e]

[a] *Department of Primary Care and Public Health, Imperial College London, UK*
[b] *Institute of Digital Healthcare, WMG, University of Warwick, Coventry, UK*
[c] *Institute of Informatics, Wroclaw University of Technology, Poland*
[d] *HRB Centre for Primary Care Research, Royal College of Surgeons of Ireland, Dublin, Ireland*
[e] *Dept of Primary Care and Public Health Sciences, King's College London, London, UK*

Abstract

The Learning Healthcare System (LHCS) refers to the close coupling of clinical research and the translation of research into practice in a cycle of continuous improvement. This vision permeates multiple domains, clinical as well as technical, and its realization is dependent on establishing standardized, secure, and traceable flows of data between these domains to maximize the research and clinical benefits. The figure below shows how different types of knowledge can interact in a single software landscape.



This panel presents the model-driven software architecture designed in the TRANSFoRm project (www.transformproject.eu), a large EU FP7 Integrated Project to develop a digital infrastructure for the LHCS in European Primary Care. The discussion will cover various components of the system, comparing them with similar tools in USA and Europe, and analyze how our modular approach supports collaboration with related efforts. Four presentations will cover:

- *Overview of the components of TRANSFoRm model-driven software architecture*
- *The TRANSFoRm software configuration for conducting epidemiological studies from primary care data sources*
- *The TRANSFoRm software configuration for electronic data collection in clinical trials*
- *The TRANSFoRm software configuration for diagnostic support*

A discussion will provide an overview of other LHCS software work in the USA and Europe, with opportunities for collaboration and international standards development. We shall also detail the software engineering practices we used in developing the architecture, and dealing with highly heterogeneous domain models and established software standards associated with the domains.

Keywords: Software Engineering, Translational Research, Randomized Controlled Trials, Decision Support

Panel description

The widespread adoption of electronic medical records and better understanding of the informatics requirements of both clinical research and transactional knowledge provide an opportunity to accelerate both research and knowledge translation, via a digital infrastructure.

Expressly coupling research methods with methods for the rapid adoption of research findings is termed 'The Learning Healthcare System'. This panel presents three core components of this infrastructure, as developed for the EU FP7 Project TRANSFoRm, alongside a discussion relating this work to international perspectives, particularly in the US.

Vasa Curcin PhD

Imperial College London, London, UK

Overview: TRANSFoRm model-driven software architecture

The EU FP7 TRANSFoRm project will produce a digital infrastructure to support translational research, by facilitating the reuse of routinely collected primary care data in different types of clinical research and providing mechanisms for reusing data collected during trials in clinical data repositories. Associated with this is the decision support component that uses the same infrastructure to provide diagnostic recommendations at the point of care, based on the knowledge extracted from routinely collected data. In this way, TRANSFoRm provides a software platform for the Learning Healthcare System paradigm, in which each component of the health system is treated as both a producer and consumer of knowledge.

The components of the TRANSFoRm digital infrastructure consist of software tools and underlying models, which are easily combined in software stacks, akin to the LAMP model[1]. In such way, the users of TRANSFoRm technology will be able to pick and choose the components they require, possibly from multiple providers, safe in the knowledge that the resulting composite system (aka stack) will seamlessly work together. Three clinical use cases demonstrate three possible approaches to the use of the infrastructure. The diabetes use case requires the TRANSFoRm technology to retrieve genotypic and phenotypic data from multiple data sources for an epidemiological study. The Gastroesophagal Reflux Disease use case utilizes the same technology to conduct a real clinical trial: recruit patients for the trial, based on their electronic health records, issue electronic Case Report Forms and Patient Reported Outcome Measures, and collect the results that are then, together with some of the care data, fed back to the researcher. The diagnostic decision support service for abdominal pain, chest pain and dyspnea uses the same core elements to manage an evidence base and associated services required by a clinician-facing software tool.

The key novel contribution of TRANSFoRm is its fully model-based approach. The models form the backbone of the tools and ensure their interoperability both on the conceptual level, and for concrete data exchange tasks. This facilitates the construction of a semantically aware provenance trace that ensures the full auditability of the software architecture.

Theodoros Arvanitis DPhil

University of Warwick, UK

Using TRANSFoRm for Epidemiological Studies

The Query and Data Extraction Workbench is the key component in the TRANSFoRm Epidemiological Study configuration. The main aim of this tool is to automatically identify 'prevalent cases' for research, where the searches will report back counts of eligible subjects in the EHRs, flagging the subjects for recruitment and consent by the local clinical care team, in full compliance with data protection legislation and best practice. In the subsequent step, a data extraction request is sent, identifying data elements required to retrieve the relevant clinical information for the diabetes epidemiological study. The Query and Date Extraction Workbench conforms to the Clinical Research Information Model (CRIM), the underlying information model for TRANSFoRm, which covers Good Clinical Practice (GCP) compliant randomized clinical trials and data collection from electronic Case Report Forms (eCRFs), EHR and web questionnaires, as well as aggregated primary care databases.

The Query and Data Extraction Workbench provides a central interface for clinical researchers to collaboratively create clinical studies, define eligibility criteria, run distributed queries, monitor query progress and report on query results. These tasks form an important part of the clinical trial recruitment process as they provide feasibility indicators on the complexity of eligibility criteria and from where participants can be recruited. Researchers often need to refine study protocols, through several iterations, to adjust the associated eligibility criteria. Individual researcher decisions during this process are preserved through the provenance infrastructure supporting the tool, by its close integration to the Query and Data Extraction Workbench.

Piotr Brodka PhD

Wroclaw University of Technology, Poland

TRANSFoRm technology for clinical trials

The Electronic Data Collection tool developed in TRANSFoRm is used to automate the Electronic Case Report Form data collection from clinicians and patients during clinical trials and provide bidirectional integration with EHR systems, whereby trial information can be retrieved from the EHR system, and the suitable part of collected data passed back to the EHR. The study definition model for the trials is built around the CDISC Operational Data Model (ODM). Study definition contains all the required metadata, including eCRFs, Patient Reported Outcome Measures forms (PROM), patient eligibility criteria, and study timeline. When the patient is enrolled to study all avail-

[1] Originally standing for Linux-Apache-MySQL-PHP, term now denotes a family of interoperable open-source technologies.

able information about study participant is populated from the EHR integrated with the system, with the remainder filled out by the clinician using either the built-in EHR data forms or the generic TRANSFoRm web tool. Study participants can fill out the PROMs using either the mobile application (Android, iOS) or web tool.
The Data Collection Server monitors the trial, sending the reminders to patients to fill out the PROM form, alerting the clinician/ researcher if alarm symptoms/signs are reported, and storing the answers in Study Database. The tool is integrated with TRANSFoRm middleware, which provides authentication and authorization services, provenance support and semantic vocabulary services.

In addition to improving data collection through eliminating paper forms, the main advantage of the approach is that the entire trial and information workflow is integrated into a single mechanism with a single underlying model. This facilitates process audits and reduces the administrative overhead when managing a study.

Derek Corrigan MSc
Royal College of Surgeons of Ireland, Dublin Ireland.

Diagnostic support using TRANSFoRm technology

The vision of the Learning Healthcare System implemented in TRANSFoRm also addresses the important translational process of generation of actionable clinical knowledge from electronic sources of primary care data. The main output to support this is provided by a clinical evidence web service that describes the concepts and relationships required to support diagnostic decision support in three chosen clinical use cases: dyspnoea, chest pain and abdominal pain. The service implements a reusable model of clinical evidence in the form of a clinical evidence ontology. Content population of the ontology is supported by a data mining module that derives quantified rule based associations from primary care coded EHR data for update into the ontology service.

We describe the overall decision support architecture by showing how other reusable TRANSFoRm technologies, such as the provenance model, EHR data integration model and security authentication are deployed as part of a decision support solution. The discussion will focus on the development of the Clinical Evidence Model and associated data mining tools, along with initial design proposals for how to effectively deploy this information using a diagnostic decision support interface integrated with a chosen EHR to provide early and late decision support during the clinical consultation.

Brendan Delaney MD
King's College London, London, UK

Discussion: Software approaches to the Learning Healthcare System

Several projects around the world are approaching aspects of the digital infrastructure for the learning healthcare system. CDISC is extending standards for the representation of clinical trial case report forms, data extracts and trial protocols, which further work on deployment tools carried out by the IHE initiative. Within knowledge translation, openCDS has provided a framework for computation of care standards, and a variety of diagnostic decision support systems have been developed over the years. The building blocks of the electronic healthcare record, terminologies, interfaces and transport tools have also been defined by ISO standards. TRANSFoRm as a project is putting all these standards to use via a common model and ontology-based approach. Even within Europe many barriers, technical, legal and organizational exist to hinder this effort and examples will be given of several recent and current initiatives. The discussion will consider whether the TRANSFoRm approach is feasible and applicable in the US and other countries and how it fosters collaboration.

All participants have agreed to take part in the panel.

This project is partially funded by the European Commission under the 7th Framework Programme, Grant Agreement Number 247787, Translational Research and Patient Safety in Europe (TRANSFoRm).

Big Data Analytics in Clinical Research

Lisa Dahm, PhD: Director, Center for Biomedical Informatics, University of California Irvine Orange, CA
Scott Duvall, PhD: Associate Director VA Informatics and Computing Infrastructure, University of Utah Salt Lake City, UT
Lewis Frey, PhD: Assistant Professor Biomedical Informatics Department, University of Utah Salt Lake City, UT
Leslie Lenert, MD, MS, FACP, FACMI: Chief Research Information Officer, Medical University of South Carolina Charleston, SC

An unsolved problem in health informatics is how to apply the past experiences of patients, stored in large-scale medical records systems, to predict the outcomes of patients and to individualize care. One approach to prediction, heretofore impractical, is rapidly finding a patient cohort “similar enough” to an index case that the health experiences and outcomes of this cohort are informative for prediction. This task is formidable because of large variability of the vast numbers of patient attributes with the added complexity of sequences of patient encounters evolving over time. Epidemiological considerations such as confounding by indication for treatment also come into play.

The panel will discuss improvements and issues with the use of big data methodologies for predictive analytics in clinical research. Big data will be characterized by volume, variety, veracity and velocity. The panel is made up of four clinical informatics researchers that are involved with the development of big data systems in healthcare. Since there is an active collaboration among the panelists on big data solutions, they will describe synergy that is achieved. The panelist will draw upon their experience with these three big data solutions and describe implications for predictive analytics in healthcare. The audience will be asked about their experience with data warehouse technology and how they see the paradigm changing with NoSQL systems. The audience also will be asked to participate by describing what big data initiatives are underway at their institutions.

Panelist Overview
Dr. Lisa Dahm, the Director of Clinical Informatics at the University of California Irvine's Medical Center (UCIMC), oversees the development and operations of Saritor, a big data solution for their Electronic Medical Record (EMR). Dr. Dahm will provide an overview of Saritor and discuss key decision made related to deploying their big data ecosystem at UCIMC. Dr. Scott Duvall is the Associate Director for the Department of Veterans Affairs’ (VA) Informatics and Computing Infrastructure (VINCI) database, which includes natural language processing (NLP) software for markup of all the VA patient records collected across the US. Dr. Duvall will describe VINCI along with the challenges of integrating and analyzing data aggregated across the United States. He will relate the solutions deployed at the VA in a traditional relational database with that of NoSQL systems deployed by the panelists. Drs. Leslie Lenert and Lewis Frey are Co-PIs on a NIH funded Clinical Personalized Pragmatic Prediction of Outcomes (Clinical3PO) big data system deployed at the VA. The Clinical3PO initiative is focused on predictive analytics within the VA using similarity matching technology. Dr. Lenert will present an overview of the Clinical3PO system and its implications for clinical care. Clinical3PO is related and integrated with VINCI through translation of the data into Clinical3PO. Dr. Frey will discuss the technology development and preliminary results from the near-term prediction algorithm within the Clinical3PO system.

Saritor, a big data solution for the EMR at UCIMC (Dr. Dahm)
Most EMR systems have not been developed to handle complex operations such as anomaly detection, machine learning, building complex algorithms or pattern set recognition. EMR systems are primarily transactional taking feeds from source systems via an interface engine. Enterprise Data Warehouses tend to be a collection of data from the EMR and various source systems in the enterprise. Enterprise Data Warehouses suffer from a latency factor of up to 24 hours. The Enterprise Data Warehouse serves clinicians, operations, quality and research retrospectively as opposed to real time. A healthcare information ecosystem, built on "Big Data" technologies, is capable of serving the needs of clinicians, operations, quality and research in real time and in one environment.

The UCIMC's legacy data of 1.2 million patients, contained in 9 million patient medical records was successfully ingested into the *Saritor* Hadoop Distributed File System. HL7 messages from all source systems, physiological monitoring data in one-minute intervals, and ventilator data in one-minute intervals and EMR generated data were ingested and stored. Algorithms for sepsis, hospital acquired conditions and 30-day readmits were built into Mahout for real time surveillance. For researchers, data visualization was provided via the drag & drop query and visualization tools. For clinicians complete inpatient care records were retrievable via a web browser.

VA's VINCI Database (Dr. Duvall)
VINCI is an environment established to facilitate research while maintaining veteran's privacy and security, and it offers the most complete collection of Electronic Health Records (EHR) from veterans. The VINCI platform combines a complete copy of all veterans EHRs with NLP software for markup, parsing and interpretation along with data management and analysis applications. Through VINCI's partnerships with the VA Corporate Data Warehouse (CDW) and the Consortium for Health Informatics Research (CHIR), a wide range of data sets are available for research, including, but not limited to, patient care encounter/visit records; UMLS tagging codes derived from NLP for both clinical and administrative data elements; vital signs; prescription data; laboratory data; demographic information; text notes such as on demographics, progress, discharge, and radiology; discharge summaries; CDW extractions from Veterans Health Information Systems and Technology (VistA); Veterans Health Administration (VHA) Medical SAS. VINCI staff members provide a range of support services such as setting up isolated virtual machines for development activities along with database and systems architects to optimize database queries by indexing commonly used terms and elements of data. Taking advantage of big data tools, like distributed computing, would allow seamless querying across different data sources and datatypes, and the ability to deliver subsets of data on demand. A subset being examined with big data tools consists of 2.5 million patients in VINCI with type 2 diabetes, defined as having at least one ICD9 code of 250.x0 or 250.x2, 3.4% female and 47.9% with age greater than 65 at first diagnosis.

Clinical Personalized Pragmatic Prediction of Outcomes (Clinical3PO) Big Data System (Drs. Lenert & Frey)
Extant findings support empiric prediction of outcomes from similar patients is possible. For example, McCormick and colleagues (McCormick, Rudin et al. 2011) use Bayesian prediction rules from patient data. Neuvirth and colleagues (Neuvirth, Ozery-Flato et al. 2011) describe the application of clinical data prior to an index event to empirically predict response to treatment in diabetes. The goal of Clincal3PO is to develop approaches that apply "big data" methodologies, including Hadoop and Accumulo, to store "medical log" files. The content of these "logs" will be processed in combination with strategies for conceptual markup of events and matching of event streams, to rapidly retrieve and identify patients that are sufficiently similar to an index

case to be able to make clinical personalized pragmatic predictions of outcomes. The complete Clinical3PO systems will use the variability in next possible steps for diagnosis or treatment, predicts the change in a specified parameter (e.g., blood pressure) over a short period of time, and assesses probability of positive results for diagnostic tests, or the probability of an adverse event.

The objectives are to (1) create a modular test bed that uses a "big data" systems architecture to support research in rapid individualized prediction of outcomes from large clinical repositories and (2) to explore various approaches to making "pragmatic" near-term predictions of outcomes from clinical data. Using the VA's VINCI system, Drs. Lenert and Frey are exploring two synergistic strategies for rapidly finding a cohort of patients that are similar enough to an index patient to predict near-term treatment response and/or adverse effects in an elastic cloud environment: (1) temporal, based on similarity in the sequence of clinical events (including epidemiology abstractions) experienced by the patient, and (2) clinical or biological, based on reported history, clinical features, genomics and laboratory testing. This also includes epidemiology similarity based on certain critical events in a patient's history (for example, past exposure to a drug or chemical, or the current presence of a set of symptoms or physiological derangements).

The focus is on predicting treatment outcomes in patients with type 2 diabetes, a growing epidemic. The VINCI database has 2.5 million patients with at least one ICD-9 code representing a diagnosis of type 2 diabetes. Drs. Frey and Lenert extend the notion of patient cohorts to including modeling of the response to treatment over time and understanding what variations in the trajectory of response are important. The ability to compare outcomes across different patients' cohorts will support the development of novel quality measures and cost control strategies, addressing the critical issue of determining "value" in healthcare.

At the limit, a perfectly matched cohort would show the distribution of outcomes subject to random variation. However, because no cohort will be perfectly matched to a patient, they hypothesize that there is an optimal degree of similarity for prediction that can be approximated by systematic exploration of the space of similarity matching across time and concepts in the medical record. Biological sequence alignment algorithms are useful in understanding which variations in sequences are predictive of outcomes and best define a cohort. Alignment of the sequence of events experienced by patients may be an important factor in creating unbiased cohorts for prediction. Extending the sequence of events requires better, faster and more flexible technologies for assessment of similarity, and it is only possible with access to very large databases of patients, where one can compare an index case to millions of other patients to find the similar cohort.

Summary

Implementing big data solutions holds many opportunities. Dr. Dahm's initial findings demonstrate that the Hadoop ecosystem is well suited for the ingestion, storage and retrieval of both legacy EMR data and runtime EMR data. Minimal programming is required to process legacy data and the processing of runtime EMR data requires the cloning of existing interfaces. Dr. Duvall is involved in combining NoSQL systems like Hadoop with traditional data warehouse databases to create powerful new analytic strategies. Drs. Lenert and Frey are developing strategies for producing cohorts of nearest neighbors for prediction of near-term outcomes designed to improve and personalize patient care. Hadoop and other big data systems provide an ecosystem that is affordable, scalable and highly available, while allowing clinical research and clinical practice to coexist in the same system.

All four panelists have agreed to participate in the panel.

Implementation of Cloud Service vs. Locally-Produced Clinical Decision Support to Assess Traumatic Brain Injury Risk in Children: A Multi-Center Study

Organizer: Marilyn D. Paterno, MBI[4,5]

Panel Moderator: Peter S. Dayan MD, MSc[1]

Panelists: Eric Tham MD, MS[2,3], Howard S. Goldberg MD[4,5], Robert Grundmeier MD[6], Marilyn D. Paterno, MBI[4,5] , for the Traumatic Brain Injury Study Group of the Pediatric Emergency Care Applied Research Network (PECARN)

Co-Principal Investigators: Nathan Kuppermann, MD, MPH[7] (not a panel speaker), Peter S. Dayan MD, MSc[1]

[1]Columbia University College of Physicians and Surgeons, New York, NY, [2]University of Colorado School of Medicine, Aurora, CO, [3]Children's Hospital Colorado, Aurora, CO, [4]Partners HealthCare System, [5]Brigham and Women's Hospital and Harvard Medical School, Boston, MA, [6]The Children's Hospital of Philadelphia, Philadelphia, PA, [7]Nationwide Children's Hospital, Columbus, OH, [7]University of California Davis School of Medicine, Sacramento, CA

Abstract

The overall goal of this multi-center study is to decrease inappropriate use of cranial CT for children with minor blunt head trauma (BHT) by creating a generalizable model to translate evidence into clinical practice. Participating sites used either a web-based, platform-independent Clinical Decision Support (CDS) Service provided by the Enterprise Clinical Rules Service (ECRS) team at Partners HealthCare System (PHS) or locally produced CDS (i.e. using the EHR's CDS rules engine) developed at a central site and exported to sites selecting the local CDS option. This panel will describe the process of creating specific, computable knowledge from evidence for use across multiple institutions. A key contribution to the field of generalizable computer decision support that we provide is our experience using the same decision support content in disparate CDS systems. Our learning goals for this panel are three-fold: [a] to understand the processes needed to provide CDS for multiple sites both within a local EHR internal rules engine and from an external, remote, cloud-based CDS service; [b] to consider the pros and cons of each approach; and [c] to understand how best to provide shareable, reusable, scalable, and maintainable CDS. We will provide initial findings from implementation from two sites in our assessment. All panelists are key participants in this research, and each brings specific expertise in his/her presentation area.

Description

We are engaged in a multi-center research study in which we implement and evaluate the effectiveness of two clinical prediction rules that assess the risk of clinically-important traumatic brain injuries (ciTBI) in children younger than 2 years and 2-18 years after minor blunt head trauma (BHT). The prediction rules were derived and validated by the Pediatric Emergency Care Applied Research Network (PECARN) and published in 2009[1].

The overall goal of our implementation study is to decrease unnecessary cranial CT use for children with minor BHT. We hypothesize that an active computer-based clinical decision support (CDS) strategy to implement the PECARN TBI rules compared to passive diffusion (the standard strategy) will safely decrease the inappropriate use of CT in children with minor BHT who are at very low risk for ciTBI. Integration of these prediction rules into an EHR requires the creation of CDS to execute the rules and return information to the emergency department clinician who must decide in a timely fashion whether or not to obtain a CT scan. We focus this panel on how to accomplish these tasks, which include:

a) translating clinical knowledge (the prediction rules) into useable CDS content that can be shared among sites,
b) creating CDS logic from the content,

c) developing executable CDS rules,
d) testing the implementation to ensure consistent and accurate results, and
e) assessing our experience, with lessons learned.

The CDS has been implemented in the local EHRs at participating sites. We offered each participating site one of two CDS methodologies for their use: [1] a platform-independent CDS cloud service provided by the Enterprise Clinical Rules Service (ECRS)[2] team at Partners HealthCare System (PHS), or [2] locally provided CDS (i.e. using the EHR's CDS rules engine) developed at one site and exported to sites that chose to use that option. Two sites elected to use cloud services; local CDS is used as a fail-over and runs simultaneously at these sites . Initial results from one site demonstrate successful responses from the web service for 98% of all CDS events, 83% of which were as fast as or faster than that from the local CDS option. Creating executable CDS for each method required different approaches, which we will describe, and a thorough testing strategy to ensure consistent results regardless of the method used at each site. We anticipate that the use of CDS cloud services can provide a generalizable, platform-independent approach to the transmission and implementation of prediction rules and treatment guidelines; however, it is important to discuss the differences between the development and execution of the cloud model and the locally provided CDS model, and to understand the benefits and risks of each approach.

The panel will be moderated by the study PI and will include four expert panelists, each of whom led a specific aspect of the work. They will discuss in detail these necessary components for producing both types of CDS:

- Identifying the appropriate population at risk
- Developing a BHT data collection template for capturing patient-specific risk factors necessary to assess the risk of ciTBI for children who present to emergency departments, and a testing strategy to ensure accurate and consistent results regardless of implementation type
- Creating a specification for producing a consistent set of logic statements to determine risk of ciTBI within each CDS system
- Designing and building decision logic within the local EHR system
- Designing and building decision logic for the cloud service
- Creating software to connect the cloud service to the local EHRs (i.e., preparing the BHT template data for transmission to the cloud service, calling the service, formatting the result, and returning it to the local EHR)
- Installing, testing, implementing, and presenting results from the CDS to clinicians at the participating sites
- Assessing our experience, including early results, with lessons learned, including the benefits and risks of each CDS methodology

The panel is organized as follows:

Dr. Dayan will introduce the panel and provide a brief overview of the problem and research questions; specifically he will describe the clinical conundrum and the balanced risk of missed injuries vs. overuse of CT, comment on the challenge to influence clinician decision making, and describe the necessary collaboration between multiple disciplines to complete this project. Following the presentations, he will moderate an open discussion among panelists and all interested members of the audience, to consider questions related to the presentations and the learning goals.

Dr. Tham will describe the effort required to produce the BHT data collection template tool for identifying the appropriate population at risk and capturing patient-specific risk factors, and to build logic statements in the local EHR system from the specification. He will discuss the challenges of creating complicated decision rules in a vendor's EHR and of exporting/importing locally produced logic across multiple sites.

Dr. Goldberg will discuss the process of creating the semi-structured specification from the risk tables derived from the 2009 study, which was used for producing consistent logic statements by both CDS methodologies. He will also describe the design of the decision logic used by the cloud service and discuss building the cloud service itself, including legal and performance concerns.

Dr. Grundmeier will present the requirements for connecting the cloud service to the local EHRs, including describing the middleware created to map the BHT template data for transmission to the cloud service, call the service, and format and return the results to the EHR. He will also discuss security concerns.

Ms. Paterno will present the testing strategy used to ensure accurate and consistent implementation of the prediction rules regardless of CDS methodology used, describe the cloud service implementation at two sites, which included installing and testing all

components and designing the presentation of results using local EHR tools, and discuss the challenges presented at each site. Additionally, she will present results of these efforts, including initial findings following implementation, lessons learned through the process, a comparison of the two methods utilized, and the pros and cons of each CDS approach.

We will hold all questions until the end, at which time Dr. Dayan will moderate the open discussion described above.

Time	Speaker	Topic
5 min	Dayan	Overview of the Research
15 min	Tham	Data Entry Tool, Local Rules
15 min	Goldberg	Semi-Structured Specification, Cloud Service Rules
15 min	Grundmeier	Connections between Site and Web service, Security
15 min	Paterno	Testing Strategy, Implementation Results and Assessment
25 min	Dayan	Open Discussion

Participation Statement

All proposed panelists are aware of this panel submission and have agreed to participate in the panel if the proposal is accepted.

Study Authorship

Marilyn D. Paterno MBI[1,2,a,b]
Eric Tham MD, MS[3,6,b]
Howard S. Goldberg MD[1,2,b]
Robert Grundmeier MD[4,b]
Jeffrey Hoffman MD[5]
Marguerite Swietlik MSN, CRNP[3]
Molly Schaeffer MS[1]
Deepika Pabbathi MSc(IS)[1]
Vickie Schum RN, BSN[7]
Allen Cole RN, BSN[8]
Sara Deakyne MPH[4]
Beatriz A. Rocha MD, PhD[2]
Dustin Ballard MD, MBE[9]
Nathan Kuppermann MD, MPH[8,10,d]
Peter S. Dayan MD, MSc[11.c,d]
for the Pediatric Emergency Care Applied Research Network (PECARN) and Clinical Research in Emergency Services & Treatments (CREST) Network

[1]Partners HealthCare System, Boston, MA, [2]Brigham and Women's Hospital and Harvard Medical School, Boston, MA, [3]Children's Hospital Colorado, Aurora, CO, [4]The Children's Hospital of Philadelphia, Philadelphia, PA, [5]Nationwide Children's Hospital, Columbus, OH, [6]University of Colorado School of Medicine, Aurora, CO, [7]Cincinnati Children's Hospital Medical Center, Cincinnati, OH, [8]University of California, Davis Medical Center, Sacramento, CA, [9]Kaiser Permanente, San Rafael Medical Center, San Rafael, CA, [10]University of California, Davis School of Medicine, Sacramento, CA, [11]Columbia University College of Physicians and Surgeons, New York, NY

[a]Panel Organizer, [b]Panelist, [c]Panel Moderator, [d]Study PI

Acknowledgement

American Recovery and Reinvestment Act-Office of the Secretary (ARRA OS): Grant #S02MC19289-01-00. PECARN is supported by the Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB), Emergency Medical Services for Children (EMSC) Program through the following grants: U03MC00001, U03MC00003, U03MC00006, U03MC00007, U03MC00008, U03MC22684, and U03MC22685.

References

1. Kuppermann N, Holmes JF, Dayan PS, et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. Lancet. 2009 Oct 3;**374**(9696):1160-70.
2. Goldberg HS, Paterno MD, Rocha BH, et al. A highly scalable, interoperable clinical decision support service. J Am Med Inform Assoc. 2013 Jul 4.

Implementing a Clinical Research Management System: One Institution's Successful Approach Following Previous Failures

Thomas R. Campion, Jr., PhD[1,2,3,4], Vanessa L. Blau, BA[1], Scott W. Brown, MA[1], Daniel Izcovich, BA[1], Curtis L. Cole, MD[1,2,5]

[1]Information Technologies and Services Department, Weill Cornell Medical College, New York, NY, [2]Center for Healthcare Informatics and Policy, Weill Cornell Medical College, New York, NY, [3]Department of Public Health, Weill Cornell Medical College, New York, NY, [4]Department of Pediatrics, Weill Cornell Medical College, New York, NY, [5]Department of Medicine, Weill Cornell Medical College, New York, NY,

Abstract

Clinical research management systems (CRMSs) can facilitate research billing compliance and clinician awareness of study activities when integrated with practice management and electronic health record systems. However, adoption of CRMSs remains low, and optimal approaches to implementation are unknown. This case report describes one institution's successful approach to organization, technology, and workflow for CRMS implementation following previous failures. Critical factors for CRMS success included organizational commitment to clinical research, a dedicated research information technology unit, integration of research data across disparate systems, and centralized system usage workflows. In contrast, previous failed approaches at the institution lacked a mandate and mechanism for change, received support as a business rather than research activity, maintained data in separate systems, and relied on inconsistent distributed system usage workflows. To our knowledge, this case report is the first to describe CRMS implementation success and failures, which can assist practitioners and academic evaluators.

Introduction

Institutions increasingly rely on electronic systems for the conduct and administration of clinical care and research. For clinical care, electronic health record (EHR) systems enable documenting the practice of medicine, nursing, and ancillary services while practice management (PM) systems facilitate patient registration, scheduling, and billing. For clinical research, clinical data warehouses and electronic data capture tools enable reuse of EHR data and prospective data collection while electronic institutional review board (eIRB) and clinical research management systems (CRMSs), also known as clinical trial management systems (CTMSs), facilitate administrative functions such as human subjects protection and clinical research billing compliance.

Clinical research billing compliance involves determining which items and services in a study protocol reflect conventional care and are therefore billable to insurers versus those that are specific to a research study and are only billable to a research sponsor[1]. Accurate clinical research billing is necessary to ensure reimbursement for research procedures and also to prevent overbilling to, and overpayment from, insurers. Billing compliance failure, particularly with Medicare in the United States, can result in considerable financial penalties and negative publicity.

CRMSs facilitate billing compliance by enabling clinical researchers to define research protocols with basic characteristics such as title and sponsor, perform prospective reimbursement analysis of procedures in a protocol, generate protocol budgets for internal analysis and negotiation with sponsors, and manage enrollment of subjects in protocols. Shared with PM and EHR systems, CRMS data enables institutions to schedule and identify patient visits with research services, distinguish research charges from standard of care, and make clinicians aware that patients are receiving experimental care.

Adoption of EHR and PM systems continues to increase[2] along with systems for the conduct of clinical research such as i2b2[3] and REDCap[4]. However, adoption of systems for the administration of clinical research, particularly CRMSs, lags in comparison[5]. Although adoption of eIRB systems is relatively common, adoption of CRMSs is low within and across institutions. Specifically, one institution reported 25% voluntary adoption of a CRMS[6] while a recent survey showed 35% adoption by academic health centers, the lowest percentage for any system surveyed for the administration and conduct of research[5].

CRMS use can streamline administrative workflows for clinical research, and understanding best practices for implementation, especially those related to people and organizational issues[7], are vital to ensure adoption and continued use of systems[8]. To our knowledge, literature describing CRMS implementation is limited to one solution

at one institution[9]. Additional description of successful and failed CRMS activities can inform practitioners at other institutions as well as academic evaluation efforts. The goal of this case report is to describe a successful approach to CRMS implementation in the context of previous failures at one institution. We consider a CRMS as a tool that facilitates research administration workflows rather than research conduct workflows such as data capture in an electronic case report form[10].

Background

Institutional environment

Weill Cornell Medical College of Cornell University (WCMC), located in New York City on the Upper East Side of Manhattan, serves a tripartite mission of clinical care, education, and research. Clinical faculty members practice in the Weill Cornell Physician Organization, an 883 physician group practice providing outpatient primary and specialty care at 22 locations citywide, and have admitting privileges at NewYork-Presbyterian Hospital (NYPH), a 2,409-bed six-facility teaching hospital. WCMC trains more than 950 resident physicians and fellows as well as over 400 medical students. To assist researchers, WCMC offers specialized services through 25 core facilities. Recent clinical and translational research initiatives include formation of a Cancer Center, Institute for Precision Medicine, and federally-funded Clinical and Translation Science Center. To support WCMC's tripartite mission, the Information Technologies and Services Department (ITS) provides comprehensive electronic infrastructure including network connectivity, software hosting, and user support among other activities. Physician Organization Information Services (POIS) works closely with ITS to leverage institutional architecture for providing clinical information systems.

WCMC and NYPH, longtime clinical affiliates and separate legal entities, increasingly collaborate for clinical and translational research. In January 2013, WCMC and NYPH established the Joint Clinical Trials Office (JCTO) to grow clinical research activities across both institutions by sharing infrastructure. With the goal of advancing patient care, basic research, and education, the JCTO streamlines the administration and conduct of clinical research through master agreements with funding sources, scientific and feasibility review of study protocols, financial management, regulatory compliance including human subjects protection, and information technology (IT).

Both WCMC and NYPH have extensive EHR, practice management, clinical data warehouse, and electronic data capture solutions. Notably, the WCMC Physician Organization has used Epic Ambulatory as its EHR since 2000 while NYPH has installed Allscripts Sunrise Clinical Manager in inpatient and emergency settings. The two institutions share electronic patient data through multiple interfaces.

Institutional CRMS requirements

WCMC's requirements for a CRMS include the ability to create and maintain protocol definitions, subject enrollments, prospective reimbursement analyses, and budgets as well as share data with PM and EHR systems. A protocol definition includes elements such as a short version of a study title, IRB protocol number, principal investigator name, fee schedule determined by the combination of protocol initiator (sponsor or investigator) and funding type (federal or non-federal), college fund account number (if applicable), and hospital research identifier (if applicable). A subject enrollment includes an IRB protocol number, patient medical record number, and enrollment status (e.g. enrolled, completed, ineligible). Prospective reimbursement analysis involves creation of a billing grid detailing the payer (e.g. standard of care, research sponsor) of each procedure for each visit in a protocol. A budget leverages a billing grid to identify costs and charges for each research procedure and other fees in a protocol. Sharing these data between CRMS and PM and EHR systems automates patient care research workflows.

Institutional CRMS failures

From 2007 to 2011, the WCMC Physician Organization Business Office (POBO), which managed the General Electric (GE) Centricity Business practice management system, pursued multiple approaches to CRMS implementation. During this time period, Epic Systems indicated it had no plans to develop a CRMS but would add functionality to existing modules as necessary to support research.

As a first step in in 2007, the POBO adapted existing GE Centricity Business functionality for clinicians managing patients with occupational health cases that were billable to corporations so that researchers could manage subjects enrolled in protocols with services billable to sponsors. This workflow required POBO analysts to manually create protocol definitions in the system and investigator teams institution-wide to access the system to update the enrollment status of subjects in protocols. A custom interface between GE and Epic informed clinicians of research patient visits. Additionally, the WCMC Office of Billing Compliance (OBC) began requiring all investigators to

submit a spreadsheet template-based prospective reimbursement analysis for all protocols submitted to the IRB regardless of the protocol having clinical procedures.

The case-based subject enrollment approach was intended to be temporary, as in March 2008 the institution began deploying GE Patient Protocol Manager (PPM) after years of development in partnership with the vendor and two other academic medical centers. PPM featured dedicated protocol definition, subject enrollment, and billing grid creation for prospective reimbursement analysis and budgeting as well as direct integration with GE billing and a custom interface to Epic informing clinicians of research activities. To implement PPM, a dedicated CRMS team conducted customized training and workflow development in individual departments for investigators and their teams. However, after system go-live in seven of 22 clinical areas, WCMC halted PPM implementation in November 2010 due to user dissatisfaction with inflexible system features, including adding procedures to billing grids and adjusting incorrectly specified procedures in downstream billing. At this point, all departments reverted to the previous case- and spreadsheet-based legacy approach, and the CRMS team began hosting vendor product demonstrations to identify a replacement system.

In March 2011, internal audit revealed that investigators were inconsistently using the legacy system for subject enrollment. Investigators cited frequent turnover of research coordinators, who often worked on protocols during their "gap year" before attending medical school, as a point of failure in the subject enrollment process. Audit findings resulted in a policy change that required investigators and their teams to use the Jira bug tracker system to inform central administration staff to create and update subject enrollment data in GE Centricity Business. Multiple attempts at CRMS implementation by the WCMC POBO failed over the course of almost four years.

Methods

In response to previous implementation failures, WCMC adjusted its approach to organization, technology, and workflow for CRMS.

Organization

In April 2011, management reassigned the CRMS team, technology, and business processes from the POBO to ITS, the college-wide IT department, because CRMS was considered a research activity rather than a practice management activity. In May 2011, WCMC licensed StudyManager Reveal, now known as Merge CTMS Investigator, for use as the college-wide CRMS. In addition to providing a solution within budget, the vendor agreed to partner with WCMC in developing a CRMS tailored to the needs of academic medical centers. Shortly thereafter, WCMC PO announced plans to migrate from GE Centricity Business to Epic Practice Management—specifically Prelude for registration, Cadence for scheduling, and Resolute for billing—to more closely integrate PM and EHR activities.

Expanded systems portfolio

After the restructuring of WCMC central research administration in March 2012, the CRMS's team portfolio expanded to include systems for managing animal protection, laboratory, human subjects protection, funding awards, and conflicts of interest. Recognizing the need for dedicated research IT resources, WCMC established Research Administration Computing (RAC) within ITS to consolidate more than twenty systems and processes for the administration of basic and clinical research. RAC consisted of the existing CRMS manager and team plus personnel responsible for other administrative systems. Notably, the reporting staff for CRMS inherited reporting responsibilities for eIRB, broadening the RAC team's understanding of available administrative data for research.

Process owner and data steward

In January 2013, the WCMC-NYPH Joint Clinical Trials Office commenced operations and assumed control of business processes for the CRMS and other research administration systems while RAC took the role of data steward. This change followed more than a year of the CRMS implementation team attempting to establish business processes for system use at the institution without a clear mandate for change or mechanism of enforcement. At about the same time, POIS established June 2013 as the go-live date for the new practice management system, which necessitated simultaneous institution-wide CRMS go-live to maintain subject enrollments linked to research visits for scheduling and billing. In preparation for go-live, JCTO and RAC established weekly meetings to create and operationalize policy for CRMS implementation and ongoing use that are described below.

Technology

The CRMS vendor delivered a slightly customized version of its software while the RAC team addressed data quality issues prior to migrating records from the legacy application and developed middleware services for transmitting data across systems.

CRMS vendor system

Standard system configuration enabled creation of protocol definitions, subject enrollments, and billing grids. Per terms of the contract, the CRMS vendor developed custom interfaces between its product and WCMC's existing LDAP authentication service for user login and Epic patient index for management of study subject demographics. Additionally, the vendor delivered custom formatting of reports for budgets and prospective reimbursement analysis.

Legacy system migration

In preparing to migrate protocol and subject enrollment data from GE to the new CRMS, RAC encountered inconsistent and incomplete records. For example, the legacy application identified each protocol's principal investigator with free text such as "Smith, John," "SMITH MD, JOHN," "6SMITH MD, JOHN" or "SMITH PHD, JOHN" rather than with a unique value or username. To reliably determine the principal investigator of each study for importing into the new system required cross-referencing the IRB protocol number for the study from the GE database with the eIRB system. Additionally, some protocol records lacked a fee schedule, college account number, or hospital study identifier when at least two of the fields should have had values. Resolving such differences required cross-referencing a database of clinical trial agreements under RAC's authority. Furthermore, medical record numbers of some patients had been deactivated or merged, complicating efforts to identify subjects for properly associating them with protocols. Updating subject identifiers required manual chart review. In response to the data quality issues, JCTO and RAC pursued a two-part strategy: 1) merge study and subject data from the legacy system with other institutional sources to form a complete data set prior to a one-time import into the new CRMS and 2) develop automated interfaces between CRMS and trusted electronic systems as well as workflows to ensure future data quality.

RAC middleware services

Based on recent expansion of reporting responsibility, the RAC team recognized an opportunity to automate transfer of data to CRMS from eIRB for each study—title, IRB protocol number, status, approval date, expiration date, department, study personnel usernames—along with dictionaries of protocol status types, sponsors, sponsor types, and departments—to maintain synchronization across systems and prevent duplicate data entry. Motivating this decision was JCTO's acknowledgement of IRB approval, and thus eIRB system records, as an authoritative marker of clinical research activity at the institution. RAC also explored using protocol funding sources and initiators entered by investigators in eIRB to determine fee schedules, but data were captured without sufficient consistency in eIRB for automatic transfer to CRMS. This led to capturing these data in a new separate electronic workflow tool for supporting evaluation of studies for scientific merit and financial feasibility, which RAC interfaced to CRMS. Additionally, to ensure appropriate and consistent formatting of study personnel names, RAC developed an automated interface to WCMC's user security warehouse, the historical record of usernames and actual names of personnel at the institution, for mapping to usernames imported from eIRB. RAC also cross-referenced usernames of principal investigators against an index of Epic users to determine if the user had an Epic SER record, a requirement for users specified as principal investigators for protocols imported into Epic. Data transfers occurred at five-minute intervals during business hours and once nightly for personnel data.

For sending protocol and subject data from CRMS to Epic 2012, RAC and POIS created an automated service compliant with Epic's text file-based import specification. RAC and POIS used the import specification rather than the interface specification based on the Institute for Healthcare Enterprise (IHE) Retrieve Process for Execution (RPE) profile to be compatible with the local Epic installation's multiple service areas and fee schedules. Once per minute, a RAC-created service checked CRMS for new or updated protocols and subject enrollments meeting certain criteria, and saved protocol and subject data to separate files on a network volume. POIS's Microsoft Biztalk integration engine processed files on the network volume every five minutes to import data into Epic for the creation and updating of Epic's internal RSH records for protocols and LAR records for subject enrollments.

Workflow

To ensure quality of protocol and subject enrollment information, JCTO opted to centralize CRMS workflow. JCTO elected to store all protocol definitions in CRMS via automated transfer from eIRB to enable a comprehensive view of the clinical research enterprise. In contrast, the legacy system, which was manually maintained by an

analyst, contained only protocols with billable clinical services. JCTO instructed RAC to configure services to send data from CRMS to Epic only for protocols with clinical services. Additionally, through eIRB and CRMS integration, JCTO tightly coupled IRB protocol status, the institutional designation of whether a study was active and protecting human subjects, with the ability of study personnel to perform research services. If a protocol expired or was closed by the IRB, CRMS automatically received the data, creating a hard stop for subject enrollment and updating Epic.

For subject enrollment, JCTO maintained the existing Jira-based process and disabled the ability of EHR users to enroll subjects in studies using standard functionality in Epic. Although POIS disabled the ability of all users to create protocols and most users to enroll subjects in protocols in Epic, no option existed in the EHR to restrict principal investigators from creating or modifying subject enrollments for protocols in Epic.

To create prospective reimbursement analyses and budgets for each protocol using CRMS, JCTO planned a new process and hiring of new analysts with a start date in 2014. In the interim, JCTO instructed central administrative staff to transcribe clinical service status, college account number, and hospital study identifier to CRMS from spreadsheets submitted by investigators for each protocol rather than to a standalone form as in the legacy approach.

Results

As of October 2014, CRMS contained 7,843 total protocols and had sent 776 (9.8%) protocols with active status and clinical services to Epic. For protocols sent to Epic, CRMS had processed 7,624 subject enrollment statuses for 6,829 unique subjects. Additionally, five principal investigators used Epic to enroll 32 subjects in protocols in violation of institutional policy and CRMS-Epic data integrity. For 118 protocols that lacked clinical services and were not sent to Epic, CRMS contained 6,617 subject enrollment statuses for 5,693 unique subjects.

Discussion

To our knowledge, this case report is the first to describe CRMS implementation success and failures. At our institution, the formation of an organizational unit with ownership of clinical research processes and an information technology group focused on research administration enabled implementation of a CRMS integrated with disparate systems to support centralized workflow processes. In contrast, previous failed approaches at the institution addressed CRMS implementation as an add-on to the administration of clinical care rather than as a comprehensive change to the administration of clinical research.

Integration of data across disparate systems has facilitated adoption and ongoing use of systems for the conduct of clinical care.[11] Similarly, data integration for systems involved in the administration of clinical research may assist institutions in overcoming barriers to CRMS adoption and use. At our institution, data entered by investigators and research coordinators into electronic systems for the evaluation of a study's scientific merit, financial feasibility, and ethics now have downstream effects in EHR and PM systems for clinical research awareness and billing. Reuse of administrative protocol data heighten the need for data quality assurance, source system data capture design, and user training.

While the approach we describe has to date replaced a legacy system's protocol definition and subject enrollment functionality, our institution lacks CRMS-based billing grid creation at the time of this writing. However, through CTMS implementation, JCTO and RAC have created the organizational and technological infrastructure necessary to support research billing compliance, and plan to implement a new process and staffing model for centralized billing grid creation. Future study will evaluate the effect of CRMS billing grid creation and integration with the EHR and PM systems on measures of research billing compliance such as gross collection rates[12].

As institutions increasingly adopt commercial EHR systems, they are beholden to vendors for clinical research features. At the time of our CRMS implementation, Epic EHR did not permit us to restrict the ability of principal investigators to enroll subjects into protocols within Epic, which resulted in rogue subject enrollments. Although the subject enrollment process at our institution is centralized, it remains voluntary, and auditing compliance requires record of every subject's informed consent form. At present we lack a system for electronic consent, and our installation of Epic stores consent forms as scanned documents with limited metadata. We urge industry to address clinical research needs in upcoming EHR releases.

This case report addresses the paucity of literature on CRMS implementation success and failure. Other institutions may find the centralized clinical research management approach described in this report valuable for implementations while researchers may use the illustration of organization, technology, and workflow for CRMS to inform evaluation efforts.

Acknowledgements

The authors thank Stephen Johnson, Ph.D. and Rainu Kaushal, M.D., M.P.H. for conceptual feedback and support.

References

1 Boyd CE, Meade RD. Clinical trial billing compliance at academic medical centers. *Acad Med* 2007;**82**:646–53.

2 Wright A, Henkin S, Feblowitz J, *et al.* Early results of the meaningful use program for electronic health records. *N Engl J Med* 2013;**368**:779–80.

3 Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc*;**19**:181–5.

4 Harris PA, Taylor R, Thielke R, *et al.* Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;**42**:377–81.

5 Murphy SN, Dubey A, Embi PJ, *et al.* Current state of information technologies for the clinical research enterprise across academic medical centers. *Clin Transl Sci* 2012;**5**:281–4.

6 GE Healthcare Intros Clinical Research Management Solution -. http://www.informationweek.com/healthcare/admin-systems/ge-healthcare-intros-clinical-research-m/229400912 (accessed 30 Jul2013).

7 Ash JS, Anderson NR, Tarczy-Hornoch P. People and organizational issues in research systems implementation. *J Am Med Inf Assoc* 2008;**15**:283–9.

8 Lorenzi NM, Novak LL, Weiss JB, *et al.* Crossing the implementation chasm: a proposal for bold action. *J Am Med Inf Assoc* 2008;**15**:290–6.

9 Ranganathan D, Bell M, Willett D, *et al.* Creating a research and clinical care partnership through EMR and clinical research system integration. *AMIA Summits Transl Sci Proc* 2013;**2013**:209–13.

10 Campion TR, Blau VL, Brown SW, *et al.* What's in a name? Perceptions of a clinical trials management system. *AMIA Summits Transl Sci Proc* 2013;**2013**:28.

11 Stead WW, Miller RA, Musen MA, *et al.* Integration and Beyond: Linking Information from Disparate Sources and into Workflow. *J Am Med Inform Assoc* 2000;**7**:135–45.

12 Miller DD, Getsey CL. Impact of a compliance program for billing on internal medicine faculty's documentation practices and productivity. *Acad Med* 2001;**76**:266–72.

Role of Citation Tracking in Updating of Systematic Reviews

Miew Keen Choong, PhD[1], Guy Tsafnat, PhD[1]
[1]Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine, University of New South Wales, Sydney, Australia

Abstract

We proposed to use automatic citation tracking to enhance the retrieval of new evidence for updating Systematic Reviews (SR). We tested on a Cochrane review from 2003 (updated 2010) and retrieved 12 of the papers to be added (recall 85.7%). Citation tracking yields a high proportion of the required literature.

Introduction

The four basic steps in conducting SRs: retrieval, appraisal, extraction and synthesis are worthy of automation. Omissions of relevant evidence in the first step cannot be corrected in later stages and will thus adversely affect the SR's authoritative coverage of all available evidence. Thus it is important to adapt comprehensive search strategies that include different techniques and multiple databases. Citation tracking is a method of measuring the impact of research studies based upon a systematic analysis of how often a specific research study has been cited by others. The effectiveness of citation tracking for evidence retrieval for SR updating is yet unknown. We hypothesize that citation tracking will be an effective method to identify literature for updating SRs. The objective of this study is to test how well automatic citation tracking can identify relevant literature for SR updates.

Methods

Each reference in the SR to be updated was used to query Microsoft Academic Search (MAS). Bibliographic information and the list of articles in the "cited-by" section were retrieved from MAS. The "cited-by" articles were then used to recursively search MAS again for subsequent "cited-by" articles. To control the expansion of included literature, we developed a Randomized Controlled Trial (RCT) filter. The RCT filter finds papers in PubMed, and obtains the article types. Articles not labeled as Publication Type RCT are omitted. We check manually if the new studies included in the new review but not in the original SR can be found in MAS. We evaluate based on the availability in MAS. We tested our algorithm on a Cochrane review (Antibiotics for acute maxillary sinusitis). The original review[1] was published in 2003 and an update[2] was published in 2010.

Results

A total of 52 reference strings included in the original SR. Comparing the reference lists of the two versions manually identified 21 new studies included in the update where 14 were found in manual searches in MAS. For the first iteration, we found a total of 134 unique citations including 7 (recall 50%, precision 5.2%) of the 14 citations to be added. In the second iteration, we found additional 1028 unique citations including another 5 relevant citations (recall 85.7%, precision 1.2%). With RCT filtering, the total number of citations retrieved dropped slightly to 130 (precision 5.4%) for the first iteration and 832 (precision 1.4%) for the second iteration.

Discussion

This study is the first to quantify the effectiveness of citation tracking to support SR updates. Recall of >85% shows that citation tracking using a single database is a promising technique but is not yet enough to completely automate literature retrieval for SR update. Further testing is required to show if using multiple databases would improve recall and to compare with typical SR update approach. More studies are required to derive robust conclusions.

Conclusion

We have presented a study of an automatic and recursive citation tracking system for SR update. Based on our results the system can probably be used as a decision support system for SR updaters.

References

1. Williams Jr J, Aguilar C, Cornell J, Chiquette E Dolor R, Makela M, Holleman D, et al. Antibiotics for acute maxillary sinusitis. The Cochrane Library. 2003.
2. Ahovuo-Saloranta A, Borisenko OV, Kovanen N, Varonen H, Rautakorpi U-M, Williams Jr J, et al. Antibiotics for acute maxillary sinusitis. Cochrane Database Syst Rev. 2008;2(2).

Standard-based EHR-enabled applications for clinical research and patient safety: CDISC – IHE QRPH – EHR4CR & SALUS collaboration

Christel Daniel, MD, PhD[1,2], Anil Sinaci, MSc[3] ; David Ouagne, PhD[1], Eric Sadou[1]; Gunnar Declerck, PhD[1]; Dipak Kalra MD, PhD[4]; Jean Charlet, PhD[1]; Kerstin Forsberg[5]; Landen Bain[6]; Charlie Mead[7]; Sajjad Hussain, PhD[1]; Gokce B. Laleci Erturkmen, PhD[2]

[1]INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; [2]CCS SI Patient, AP-HP, Paris, France, [3]Software Research, Development and Consultancy, Ankara, Turkey; [4]University College London, UK; [5]AstraZeneca, Sweden; [6]CDISC; [7]W3C

Abstract

Integration profiles collaboratively developed by CDISC and IHE for integrating data from Electronic Health Records (EHRs) with clinical research and pharmacovigilance are limited to resolving lexical/syntactic data integration issues and do not address semantic barriers. This paper describes the collaboration between two European projects – EHR4CR and SALUS – in implementing ISO/IEC 11179-based metadata registries (MDRs) and semantically integrated cross-platform data access. A common "semantic MDR" provides a framework for bi-directional/cross-MDR mapping and federated queries are enabled using the newly-defined IHE Data Exchange (DEX) profile. In the pilot implementation, mappings for 178 EHR4CR and 199 SALUS metadata elements were persisted in the semantic MDR. The DEX profile was then used to access semantically equivalent data elements in SALUS or EHR4CR participating EHR systems. ISO/IEC 11179-based MDRs and DEX integration profile address the goal of developing pan-EU computable semantic integration of data from clinical care, clinical research, and patient safety platforms.

Keywords : Electronic Health Records, Biomedical Research, Adverse Drug Reaction Reporting Systems, Pharmacovigilance, Terminology as Topic

1 Introduction & background

Electronic Health Records (EHRs) contain a large variety of patient-centric data. A number of investigators have noted that the ability to integrate data from EHRs with that from other domains – e.g. clinical research and post-market patient safety – could provide significant value to both domain-specific and population-centric research [1,2]. Specific topics of interest include providing trial planners with a better understanding of the available cohorts [3,4,5,6] and targeted patient recruitment [7]. Others have addressed the efficiencies of "single-source data entry" at the point of clinical care [8,9]. Finally, ongoing reporting of post-market adverse drug events could be substantially improved if patient safety monitoring platforms had ongoing access to EHR data [10,11]. However, because EHRs are not designed with a primary focus on cross-patient data aggregation, data integration between EHRs and the more difficult scope of cross-domain integration, initiatives for integrating EHRs and Clinical Research or Patient Safety are often limited to non-scalable, one-off, system (or vendor)-specific efforts.

1.1 Two European projects: EHR-enabled clinical research (EHR4CR) and patient safety (SALUS)

The EHR4CR project (http://www.ehr4cr.eu/) is an IMI (Innovative Medicines Initiative) project funded by European Union's Seventh Framework Programme and by in-kind contributions from member companies of the European Federation of Pharmaceutical Industries and Associations (EFPIA). The EHR4CR project is one of the largest public-private partnerships focused on providing adaptable and scalable solutions for reusing data from hospital EHRs for Clinical Research in various diseases. Implementations have been installed at 11 pilot sites throughout five European countries (France, Germany, Poland, Switzerland and United Kingdom). Collectively, the EHRs from the pilot sites contain data from over 7,000,000 patients.

The SALUS project (http://www.salusproject.eu/) is a STREP project funded by European Commission ICT Programme, eHealth Unit. Combining the strengths of individual case safety reports with EHR data, the SALUS project is focused on creating the necessary semantic and technical interoperability infrastructure to enable efficient and effective secondary use of EHR data in support of pro-active post-market safety studies.

Table 1 summarizes the EHR4CR and SALUS use cases as one of three high-level functional categories.

Table1. Applicability IHE integration and content profiles in the EHR4CR and SALUS contexts

High-level use cases	EHR4CR use cases Clinical Research	SALUS use cases Patient Safety
A: Identification of patient cohort based on pre-defined eligibility criteria	Protocol feasibility study and patient recruitment	Population selection for post market safety studies
B: Extraction of patient-specific data for pre-populating individual forms	Case Report Form pre-population	Individual Case Safety Report form pre-population
C: Extraction of patient-specific data for feeding a research database.		Retrospective observational study in pharmacovigilance and pharmacogenomics

1.2 Semantic interoperability

One of the main challenges in integrating cross-domain data is semantic alignment of data collected in disparate contexts by different systems. Conceptual frameworks often base solutions on the existence of a common model of "shared semantics." Common models must be based on the adoption and integration of multiple standards that themselves must be consistent, coherent, and cross-compatible [12,13,14]. Unfortunately, standards in clinical care, clinical trials, and patient safety monitoring have often been developed through parallel – and therefore somewhat inconsistent – efforts.

In the domain of **patient care**, efforts have focused on specifying both the syntax and the semantics of clinical information. The HL7 Reference Information Model (RIM) and EN 13606 standards define the semantics of patient care data and clearly demonstrate the need for "layers of semantic expressiveness" including: i) generic reference information models of concepts and relationships (e.g. CEN/ISO 13606, openEHR Reference Model, or HL7 RIM) each capable of binding terms from terminology models (e.g. SNOMED, LOINC, etc.) and associated with a data type models such as ISO 21090; and ii) more detailed models (e.g. CEN/ISO 13606 or openEHR Archetypes/Templates, or HL7 Detailed Clinical Models (DCMs), that instantiate generic reference models (e.g. HL7's Clinical Document Architecture (CDA) meta-standard and the derived Continuity of Care Document (CCD)) [15,16]. In the domain of **clinical research**, the Clinical Data Information Standards Committee (CDISC) non-profit organization has developed a number of standards for study design including (Study Design Model (SDM)[17], study data collection (Operational Data Model (ODM)) [18], study data analysis (Analysis Data Model (ADaM)), and submission to the regulatory bodies (Study Data Tabulation Model (SDTM)). Historically, CDISC standards were not defined using the "semantic layers" described for clinical care. However, in 2004, CDISC, HL7, the National Cancer Institute (NCI), CDISC, HL7, and the FDA began the development of the Biomedical Research Integrated Domain Group (BRIDG) model containing the layered representations of the semantics of regulated clinical research data and CDISC standards can now be represented as BRIDG constructs [19]. In the domain of **patient safety monitoring**, the Individual Case Safety Report (ICSR) was developed using HL7 RIM-based constructs and is therefore defined using a layered approach [20].

In the context of interoperability between clinical care, clinical research, and patient safety monitoring systems, the term of **metadata** (literally "data about data") is used to distinguish "data collection structures" from "subject-level responses" that populate those structures, i.e. instance-level. Metadata should be described using well-defined metadata schema so as to represent the semantics of the instance data and will include concepts and relationships as well as bindings to terminologies, controlled vocabularies, taxonomies, etc. Metadata scheme may be expressed in a number of different programming languages e.g. HTML, XML, UML, RDF, etc. The core international standard used to define metadata is **ISO/IEC 11179**. This standard provides the definition of a "data element" registry, describing disembodied data elements. It is important to note that ISO/IEC 11179 covers just the definition of elements and does not dictate the persistence structures or retrieval strategies. In the healthcare domain, another ISO standard – **ISO 21090** – plays a key role in the ISO/IEC 11179-based data element definitions since it provides the appropriate formal representation of the data type for Data Element Concept and of any type of the Value Domain data type. ISO 21090 especially provides a formal of the coded data types and addresses the binding with terminological systems.

Achieving broad-based, scalable and computable semantic interoperability across multiple domains requires the integration of multiple standards. Integrating the Healthcare Enterprise (IHE) (http://www.ihe.net) has emerged as the organization addressing this need. "Real-world usage scenarios" that can be instantiated using existing standards are published by IHE as Integration Profiles. Each profile defines a series of "transactions" which specify how existing standards should be applied to meet the overarching business goal. A set of integration and content profiles developed by several IHE committees - Quality, Research, and Public Health (QRPH), Patient Care Coordination

(PCC) and Information Technology Infrastructure (ITI) domains - collectively address syntactic issues of cross-vendor interoperability and are applicable in the EHR4CR and SALUS contexts [21-22].

1.3 Objective: EHR4CR-to-SALUS Semantic Interoperability

Each project develops solutions to achieve scalable, generalizable, cross-platform computable semantic interoperability. Our hypothesis is that cross-platform queries are achievable by implementing an additional "layer" of metadata defining mappings between each project's metadata. Our goal is to develop a semantic MDR persisting mappings between project-specific metadata and to implement the newly defined IHE Data Exchange (DEX) profile for accessing the common metadata and enable achieve cross-project – i.e. EHR4CR-to-SALUS – interoperability [23].

2 Methods

Independently designed and implemented, project-specific metadata models were developed to provide robust semantic definitions of all data elements needed to address each project's respective use cases. Cross-project semantic alignment was accomplished using a "semantic MDR". Cross-project, federated, semantically integrated queries were defined and managed using the DEX profile's "Retrieve Metadata" interface.

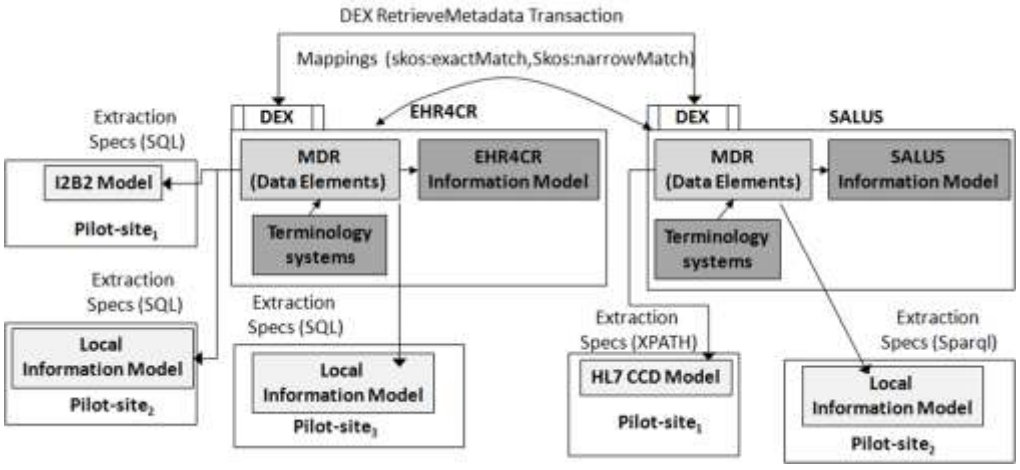
2.1 Metadata registries and semantic services

We defined the core content of the **EHR4CR** ISO/IEC 11179-based MDR as a set of data element definitions derived from HL7 RIM-based metadata models (such as HL7 CCD) bound to terminologies. This core content was enriched by specific data element definitions required to represent the semantic content in the scope of EHR4CR use cases. The concepts used in the definitions of the central data elements were mapped to corresponding local terms used in pilot sites [24]. We implemented semantic services used by the Query Builder of the Protocol Feasibility Study and Patient Recruitment modules of the EHR4CR platform (use case A). The Query Builder acting as a "metadata consumer" retrieves data elements from the MDR so that users can represent eligibility criteria as formal queries that are compliant with the common model.

The **SALUS** semantic MDR is implemented on top of a Triple Store layer that persists a relational model of ISO/IEC 11179-compliant [24] data elements and uses SKOS-based cross mappings of all metadata used by SALUS stakeholders. During the Individual Case Safety Report form population (use case B), the pre-population of the form is performed through the mappings of the data elements retrieved from the MDR.

A cross-project semantic MDR was set up in order to include mappings between SALUS and EHR4CR metadata and thereby became Sharing semantics across projects. Both EHR4CR and SALUS project maintain project-specific metadata models and the mappings of the resident data element definitions to local implementation-dependent models. To achieve cross-project interoperability, there is need for an interoperability specification to seamlessly share the definition of these data elements and the associated "extraction specifications." The Data Exchange (DEX) profile focuses on providing uniform access to shared semantics via the "Retrieve Metadata" transaction.

Figure 1. Mappings between EHR4CR and SALUS MDRs and use of IHE DEX Retrieve Metadata transaction for cross-platform query execution



3 Results

3.1 EHR4CR semantic services

The current version of the EHR4CR MDR includes 105 data elements corresponding to the semantic content of the eligibility criteria of 13 clinical trials. EHR4CR data elements are related to demographic statements (gender, birth date), diagnosis (diagnosis types (n=25 SNOMED CT codes e.g admitting diagnosis, principal diagnosis, etc) and associated value set consisting of n=12,318 ICD10 codes), findings (n=30 SNOMED CT codes and associated units or value sets), lab test results (n=34 LOINC codes and associated units or value sets), anatomic pathology observations (n=9 LOINC and associated units or value sets), procedure (value set of n=38 SNOMED CT codes), medication (substance administration)(value set consisting of n=5,655 ATC codes). Pilot sites mapped their local data structures and terminologies to the EHR4CR data elements. Using the query builder of the EHR4CR workbench, a Study Manager combines EHR4CR data elements plus logical and temporal operators to populate query-templates designed for representing formally the eligibility criteria of the clinical trial. Once a query has been constructed, it is transformed into local specific representations based on the target systems thereby identifying patients meeting specific inclusion/exclusion criteria.

3.2 SALUS semantic services

The SALUS Semantic MDR provides a federated metadata repository where machine-processable definitions of data elements across domains are shared, re-used, and semantically interlinked to enable semantic interoperability and contains services that perform all required transforms between client systems. SALUS has identified a number of local models used in the interoperating sites, e.g. OMOP CDM, ASTM/HL7 CDD, E2B, and several other proprietary models whose semantics are relevant to the supported use cases. The Individual Case Safety Report form is based on the E2B content model, and SALUS Semantic MDR has the necessary mappings between the E2B fields and SALUS data elements. On the other hand, SALUS data elements have also the necessary mappings to the ASTM/HL7 CCD content model fields and the SALUS Common Model fields. Once MDR-resident common data elements have been mapped to corresponding elements in the local content models, all data transformations can be performed using Semantic MDR-resident reasoning tools.

3.3 Sharing Data Elements and query specifications across projects

EHR4CR and SALUS data elements were cross-mapped to support cross-project semantic alignment. Although for some data element mappings – e.g. patient gender, patient birth date, discharge diagnosis, procedure – were relatively easy to map (i.e. 1-to-1), the majority required more complex mappings secondary to the underlying differences in levels-of-abstraction that exist between the two MDRs. In particular, most of the SALUS Data Elements were defined using high-level generic content models) that have been constructed using generic terms such as "Result." In contrast, EHR4CR data elements correspond to highly specific content models corresponding to specific results (e.g *Glucose [Moles/volume] in Serum or Plasma*), vital signs (e.g. *Systolic Blood Pressure)* or problems (e.g *ECOG performance status*). In the SALUS MDR, similar specificity is represented by constraining general elements using run-time, query-specific binding of value sets, a decision that was made in order to support automated terminology reasoning on data elements. Table 1 illustrates an example of mapping between two specific observations defined in the EHR4CR Data Elements and the corresponding generic SALUS Data Elements. As a result, only 5% of the SALUS Data elements were able to be directly mapped to EHR4CR data elements through skos:exactMatch. In contrast, 99% of the test set of EHR4CR data elements were mapped to SALUS data elements using skos:narrowMatch.

Table 1. Two examples of mapping between specific EHR4CR data elements and high level generic SALUS data elements

Highly specialized EHR4CR HL7 v3 Construct & Data elements	*ISO 11179 Model Construct*	Corresponding generic SALUS HL7 v3 Construct & Data elements
Observation.code 271649006-**Systolic Blood Pressure**-SNOMEDCT	*Data Element Concept*	Result.code
Observation.value Physical Quantity (PQ) Unit (e.g. if data type is PQ) = mmHg	*value_domain_datatype* *value_domain_unit_of_measurement*	Result.value ANY (PQ, CD, CO)

Observation.code 424122007- **ECOG performance status finding**-SNOMEDCT	*Data Element Concept*	Problem.condition
Observation.value Coded Ordinal (CO) 425389002-ECOG 0-SNOMEDCT 422512005-ECOG 1-SNOMEDCT 422894000-ECOG 2-SNOMEDCT 423053003-ECOG 3-SNOMEDCT 423237006-ECOG 4-SNOMEDCT 423409001-ECOG 5-SNOMEDCT	*value_domain_datatype* *Enumerated Value Domain & Permissible value/Value meaning*	Problem.condition.value ANY (PQ, CD, CO)

Once mappings between EHR4CR and SALUS MDRs are established, interoperability is managed using the DEX profile's "Retrieve Metadata" interface. For example, an EHR4CR request for patients with "Date of Birth > 1960" and "ECOG performance status >2" data and "Recent weight loss" to be performed SALUS pilot sites EHRs first locates the semantic links between the EHR4CR data elements and the corresponding SALUS data elements, then retrieves the mapping specifications of this data elements to local data base schemas as database queries. Qualified patients are returned to the metadata consumer in system-friendly form based on transformations derived for MDR mapping information.

4 Discussion & Conclusion

4.1 Contribution

In order to accomplish cross-domain semantic interoperability between domains/projects, there must be a single semantically unambiguous, processable, sharable, and technology-neutral metadata model, i.e. semantic metadata registry (MDR). The semantic MDR model should be based on a metadata definition standards such as ISO/IEC 11179. Similar construction of the semantic MDR and domain-/project-specific "local" MDRs enables efficient transformations/mappings of MDR elements to semantic MDR elements. Achieving computational semantic interoperability thus becomes a matter of defining a set of semantically unambiguous and context-neutral common metadata definitions, the universe of "shared semantics." We first utilized a set of IHE profiles that collectively address the syntactic (non-semantic) issues involved in developing computational interoperability. We then identified the need for an additional profile to address the core semantic barrier: access to semantically annotated metadata. The DEX (Data Exchange) profile provides the technical specification for access to MDRs. The DEX profile enables the specification of queries over heterogeneous systems, projects, and domains. We demonstrated an application of DEX both within and between the two projects (SALUS and EHR4CR) i.e. across three domains (patient care, clinical research, and patient safety monitoring) and multiple participating pilot operational systems.

Our experience in designing MDRs for cross-platform semantic interoperability strongly suggests the importance of two specific implementation principles: i) the utility of using ISO/IEC 11179 as the meta-meta standard around which to construct both project-specific MDRs and the common semantic MDR; and ii) the advantage of using Semantic Web (SW) tools and technologies for the representation and sharing of cross-domain semantics. In particular, the SW approach is of considerable value since it eliminates the brittle binding of semantics to syntactic schema representational models such as RDBMS tables of XSD document trees, and instead places a serialization-independent "graph" representation of semantics – both concepts and their relationships – on the table as a "first-class citizen."

4.2 Limitations and perspectives

The "devil in the details" is the construction of various mappings that are persisted at the two metadata "levels," i.e. the local MDR for each project as well as the common, cross-project semantic MDR. A comprehensive description of the issues faced by each project in defining and implementing a scalable solution for achieving and maintaining mappings between the elements stored in their MDRs and the various local models of multiple EHRs and CDWs is out of the scope of this paper. Rather, the paper focused on the details of the mapping between EHR4CR and SALUS MDRs, i.e. on the content of the semantic MDR. Generally speaking, the mapping between SALUS and EHR4CR MDRs was eased by the fact that both projects refer to similar RIM-based layered metadata representations. However, in developing this mapping, we discovered that in the current version of the SALUS MDR, metadata elements had been defined at a higher level of abstraction then those in EHR4CR. This disparity

was addressed during the mapping using skos:narrowMatch (e.g. Glucose [Moles/volume] in Serum or Plasma as part of Result and Systolic Blood Pressure as part of Vital Signs) as the mapping predicate.

Semantic alignment of concepts must still be managed via human intervention. However, when the underlying representation is based on graphs rather than tables or document trees, harmonization efforts can occur "bottom-up" rather than "top-down" without disruption of global schemas, and are always focused on "pure" semantic alignment devoid of technology bindings. The SALUS project has demonstrated the value proposition of implementing their semantic MDR using a Semantic Web-based approach. We believe that the adoption of a similar approach to the representation of both semantic standards and their derivative products will prove to be the most effective tool to realize the benefits derivable from computable semantic interoperability.

Acknowledgements: We thank the members of EHR4CR WPG2 for their contribution to the development of the EHR4CR platform. We thank Jean Charlet who provided expertise on semantic web technologies and ontology mapping.

Funding: This work was supported by the Innovative Medicines Initiative Joint Undertaking, under grant agreement no 115189, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution (IMI JU web address www.imi.europa.eu). This work was also supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT-287800, SALUS Project (Scalable, Standard based Interoperability Framework for Sustainable Proactive Post Market Safety Studies).

Authorship and Contributorship: CD, AS, GL designed the study. CD, DK and CM developed the literature review method, with input from GD. CD, GL, SH drafted the original paper with AS. GL, LB led the analysis with substantial input from CM. DO, ES, GD and DK reviewed the original paper. CM, LB, KF and DK revised the paper for submission. CD, LB and CM responded to peer reviewers.

Competing interests: None.

Provenance and peer review: Not commissioned; externally peer reviewed.

5 References

1 Prokosch H-U, Ries M, Beyer A, *et al.* IT infrastructure components to support clinical care and translational research projects in a comprehensive cancer center. *Stud Health Technol Inform* 2011;**169**:892–6.

2 Murphy SN, Dubey A, Embi PJ, *et al.* Current State of Information Technologies for the Clinical Research Enterprise across Academic Medical Centers. *Clinical and translational science* 2012;**5**:281–4.

3 Dugas M, Lange M, Müller-Tidow C, *et al.* Routine data from hospital information systems can support patient recruitment for clinical studies. *Clin Trials* 2010;**7**:183–9.

4 Weng C, Tu SW, Sim I, *et al.* Formal representation of eligibility criteria: a literature review. *J Biomed Inform* 2010;**43**:451–67.

5 Murphy EC, Ferris FL 3rd, O'Donnell WR. An electronic medical records system for clinical research and the EMR EDC interface. *Invest Ophthalmol Vis Sci* 2007;**48**:4383–9.

6 Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association: JAMIA* 2012;**19**:181–5.

7 Cuggia M, Besana P, Glasspool D. Comparing semi-automatic systems for recruitment of patients to clinical trials. *Int J Med Inform* 2011;**80**:371–88.

8 Kush R, Alschuler L, Ruggeri R, *et al.* Implementing Single Source: the STARBRITE proof-of-concept study. *J Am Med Inform Assoc* 2007;**14**:662–73.

9 El Fadly A, Rance B, Lucas N, *et al.* Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. *J Biomed Inform* 2011;**44 Suppl 1**:S94–102.

10 Bates DW, Evans RS, Murff H, *et al.* Detecting adverse events using information technology. *J Am Med Inform Assoc* 2003;**10**:115–28.

11 Linder JA, Haas JS, Iyer A, *et al.* Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. *Pharmacoepidemiol Drug Saf* 2010;**19**:1211–5.

12 European Commission. Semantic Interoperabiloity for Better Health and Safer Healthcare. 2009.http://ec.europa.eu/information_society/activities/health/docs/publications/2009/2009semantic-health-report.pdf

13 Mead CN. Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? *J Healthc Inf Manag* 2006;**20**:71–8.

14 Hammond WE, Jaffe C, Kush RD. Healthcare standards development. The value of nurturing collaboration. *J AHIMA* 2009;**80**:44–50; quiz 51–52.

15 HL7 DCM. Detailed Clinical Models - HL7Wiki. 2013.http://wiki.hl7.org/index.php?title=Detailed_Clinical_Models (accessed 29 Mar2013).

16 Goossen WTF. Using detailed clinical models to bridge the gap between clinicians and HIT. *Stud Health Technol Inform* 2008;**141**:3–10.

17 CDISC. Medical Research Study Design in XML (SDM). 2013.http://www.cdisc.org/study-trial-design (accessed 29 Mar2013).

18 CDISC. Operational Data Model (ODM) - Certification & Archiving and Interchange of Metadata. 2013.http://www.cdisc.org/odm (accessed 8 Apr2012).

19 Fridsma DB, Evans J, Hastak S, *et al.* The BRIDG project: a technical report. *J Am Med Inform Assoc* 2008;**15**:130–7.

20 ICH. E2B (R2), Electronic transmission of individual case safety reports - Message specification (ICH ICSR DTD Version 2.1), Final Version 2.3, Document Revision Feb. 1, 2001. 2011.

20 IHE QRPH. IHE Clinical Research Document (CRD). 2010.http://wiki.ihe.net/index.php?title=Clinical_Research_Data_Capture(CRD)

21 IHE QRPH. IHE QRPH Drug Safety Content (DSC). 2009.http://www.ihe.net/Technical_Framework/upload/IHE_QRPH_TF_Supplement_Drug_Safety_Content_DSC_TI_2009-08-10.pdf

23 IHE QRPH. IHE_QRPH_Data Element Exchange (DEX). http://www.ihe.net/Technical_Framework/upload/IHE_QRPH_Suppl_DEX_Rev1-0_PC_2013-06-03.pdf (accessed 16 Jun2013).

24 Ouagne D, Hussain S, Sadou E, *et al.* The Electronic Healthcare Record for Clinical Research (EHR4CR) information model and terminology. *Stud Health Technol Inform* 2012;**180**:534–8.

25 ISO/IEC 11179. Home Page for ISO/IEC 11179 Information Technology -- Metadata registries. ISO/IEC 11179, Information Technology -- Metadata registries (MDR). 2013.http://metadata-standards.org/11179/ (accessed 13 Apr2013).

Simplifying Complex Clinical Element Models to Encourage Adoption

Robert R. Freimuth, PhD*, Qian Zhu, PhD*, Jyotishman Pathak, PhD, and Christopher G. Chute, MD, DrPH
Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

Clinical Element Models (CEMs) were developed to provide a normalized form for the exchange of clinical data. The CEM specification is quite complex and specialized knowledge is required to understand and implement the models, which presents a significant barrier to investigators and study designers. To encourage the adoption of CEMs at the time of data collection and reduce the need for retrospective normalization efforts, we developed an approach that provides a simplified view of CEMs for non-experts while retaining the full semantic detail of the underlying logical models. This allows investigators to approach CEMs through generalized representations that are intended to be more intuitive than the native models, and it permits them to think conceptually about their data elements without worrying about details related to the CEM logical models and syntax. We demonstrate our approach using data elements from the Pharmacogenomics Research Network (PGRN).

Introduction

Data normalization requires transforming information into a common semantic and syntactic representation. Normalization projects are often conducted within a defined community, consortia, or network in an effort to improve data interoperability among members. These efforts are often conducted retrospectively and include a review of data dictionaries to identify groups of data elements that share common semantic meaning[1,2]. Once sufficiently similar data elements have been identified a new data element is proposed as a local "standard" for use within the defined research context. While local standards may facilitate the collection and analysis of data for a given purpose or project, the narrow scope in which they were defined often prevents their reuse in other contexts, thereby leading to the development of additional context-specific standards. The proliferation of local data standards does not address barriers to interoperability on a larger scale[3,4]. Even in cases where local standards are utilized, data sets generated by a research study often remain in a standalone data repository that is not integrated with other clinical systems because of difficulties aligning the data elements to a given EMR data model. If the data is not integrated and accessible, researchers will have limited ability to discover, mine, and reuse the data.

These issues may be addressed by Clinical Element Models (CEMs), which are hierarchical, logical models of clinical data that can be readily aligned to EMR data models[5,6]. Our previous work[2] demonstrated the use of CEMs to standardize pharmacogenomics data elements collected from the Pharmacogenomics Research Network (PGRN)[7]. In parallel, the SHARPn project demonstrated how CEMs can be used as an implementation-independent means for exchanging clinical data[8,9]. Together, these projects illustrated how CEMs can be used to avoid the creation of local data standards and to decompose precoordinated data elements into forms that better fit EMR data models.

The CEM standard is quite complex, however, and specialized knowledge and training is required to implement the models appropriately. Therefore, the same complexity that makes the standard a robust and valuable resource becomes a barrier to investigators that lack the knowledge to adopt it. The goal of this project was to address that limitation by developing a system that would enable investigators to utilize the CEM standard for standardized data representation without first requiring them to acquire specialized knowledge about the specification.

To accomplish this goal we introduced a layer of abstraction over the CEMs that allows investigators to think conceptually about their data elements without getting bogged down in implementation details. This layer of abstraction, termed "Patterns", also illustrates how "primitive" data elements within CEMs can be used to construct more complex, semantically precoordinated data elements. We demonstrate our approach using use cases from the Pharmacogenomics Research Network (PGRN), but this method is also applicable to other data standards and scientific domains[3].

Materials and Methods

The "Pattern" is at the center of our proposed approach. Patterns group together the CEM attributes that are necessary to represent a given abstract data element, which eliminates the need for an investigator to examine the

underlying models and determine themselves how to utilize the standard. The Pattern meta-model, the creation of patterns for the PGRN, and the evaluation of this approach are described below.

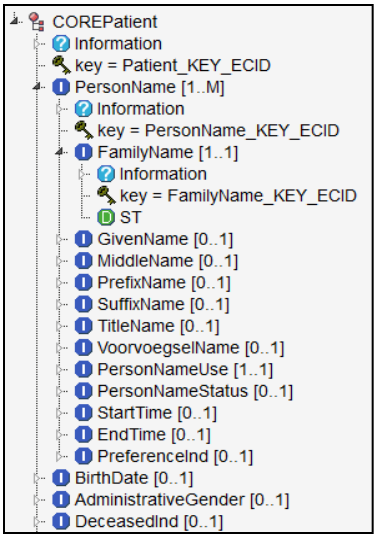


Figure 1: A portion of the SHARPn "Patient" CEM

Clinical Element Models (CEMs)

GE and Intermountain Healthcare developed a large library of CEMs, which are hierarchical, logical models of clinical data. CEMs are defined using the Constraint Definition Language (CDL)[5] and are available through the CEM Browser[4] in both CDL and XSD format. The CEMs that were used for this project were loaded into a local database to support aggregation and mapping at the attribute level. Figure 1 shows a portion of the SHARPn "Patient" CEM. The hierarchical model includes composite attributes (e.g., PersonName) that contain atomic attributes (e.g., FamilyName) and HL7 V3 datatypes (e.g., ST). Each attribute has a "key" (e.g., PersonName_KEY_ECID), a coded term that provides semantic meaning for the attribute, and cardinality (e.g., 1..M).

Pattern Meta-Model

Patterns group all of the attributes that are necessary to represent a given abstract data element into a single container, thereby eliminating the need for an investigator to examine the underlying CEMs and try to determine themselves how to represent their data using the standard models. A conceptual diagram of the Pattern meta-model is shown in Figure 2.

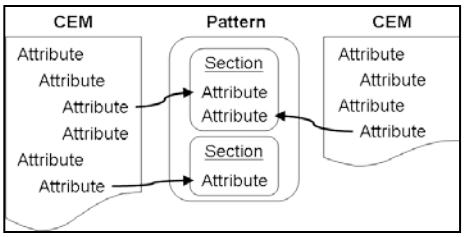


Figure 2: A conceptual diagram of the CEM Pattern meta-model. Not all components are shown (see text).

A Pattern has a name and description, and is composed of one or more Sections. Sections are logical groupings of one or more related CEM attributes. Sections have a name, definition, and an ordered list of attributes. Each attribute in a Section has a display name, datatype, and description. Sections can contain attributes from different CEMs or from different branches within a CEM hierarchy.

The ability to group attributes from structured, logical models into arbitrary sections enables the creation of conceptual abstract data elements, which can then be instantiated as study-specific data elements. It is important to note that none of the semantics of the underlying models are altered. Detailed mappings are maintained between the elements used in a Pattern and their respective source models. These mappings, as well as the source models themselves, can be hidden to simplify the information that is presented to the consumer.

Pattern Creation

The creation of a Pattern by a knowledge engineer includes three steps: the identification of an abstract data element, the identification of CEM attributes that capture the semantics of the data element, and the definition of the Pattern itself. This process helps to transform study-specific data elements into standardized representations that are more likely to be EMR-compliant and usable for secondary purposes.

The first step in creating a reusable abstract data element is to identify study-specific elements that can be generalized. It is common for investigators to define data elements based on a hypothesis or data analysis plan, as it is convenient to think about data in terms of how it will be used. Unfortunately, this often leads to the generation of study-specific, non-reusable data that is difficult to enter into an EMR due to excessive semantic pre-coordination. For example, it is common to collect age-based data in clinical trials. In previous studies we observed data elements derived from questions such as "how old was the patient when diagnosed with diabetes" and "at what age did you first experience palpitations"[1,2]. The answer to each of these questions, representing the age of the patient in years, is captured as a single integer. Both of these data elements can be generalized as the abstract data element "age at diagnosis of disease", where the specific disease is not specified until the data element is instantiated.

Once an abstract data element has been identified it must be decomposed into atomic concepts that are represented as attributes in CEMs. In our experience, this almost always results in the expansion of the number of data elements that need to be captured because data elements used in research tend to have some degree of semantic pre-coordination. For example, the abstract data element "age at diagnosis of disease" requires several attributes from

two different CEMs. The concept of "disease" is represented by a term from a controlled terminology (e.g., ICD, SNOMED-CT), and is captured in an attribute from the Disease/Disorder CEM. The concept of "age at diagnosis" is the result of a calculation that computes the difference between the date of diagnosis (from the Disease/Disorder model) and the patient's date of birth (from the Patient model) (Figure 3). Therefore, a minimum of three attributes are required for this abstract data element: disease code, date of diagnosis, and patient date of birth.

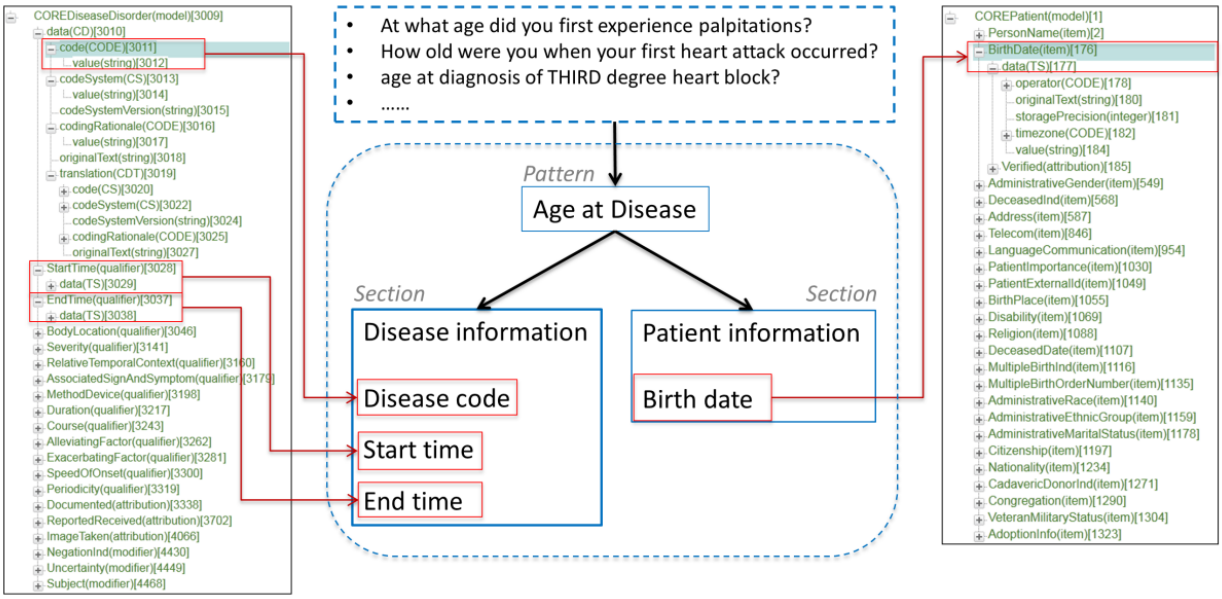


Figure 3: Identification of attributes for a Pattern. Attributes are selected from the Disease/Disorder model (left) and the Patient model (right) for the "Age at Disease" Pattern. See text for details.

A Pattern can be created once attributes have been identified. Since Patterns are intended to be intuitive, conceptual representations it may be necessary to provide user-friendly names or definitions instead of using the formal, technical ones that are part of the underlying logical model. It is important to note that the semantic meanings of the attributes are not changed and that a precise mapping is maintained between a CEM attribute and its representation within a Pattern.

Additional attributes can be added to increase the applicability of the Pattern to other, semantically similar data elements. This is consistent with the notion of creating a representation of a generalized, abstract data element. For example, the abstract data element "age at remission of disease" is closely related to "age at diagnosis of disease". The only difference between the two is whether the start date (for diagnosis) or end date (for remission) of the disease is used in the calculation of age. Therefore, it is reasonable to add this attribute to the pattern, further abstracting the pattern to simply "age at disease" (Figure 3). Note that not all attributes within a Pattern need to be used for a specific implementation, so instantiating a data element that captures age at diagnosis would not require the disease end date to be populated.

After a Pattern has been created, documentation is added to explain the purpose of the Pattern, the logical model(s) it was derived from, and how it should be instantiated for a particular implementation or research project.

User Interface

A web-based graphical user interface was developed to facilitate pattern creation (Figure 4). The figure illustrates the workflow for creating a Pattern with the web-based user interface. A new pattern is given a name and introductory text (which may include HTML tags). One or more sections are then created, which are populated with attributes that represent the semantics of the abstract data element. Attributes are selected from one or more CEMs that are stored in a local database. Attributes can be given a user-friendly display name, description, and order, which is displayed when the user views the completed pattern detail screen.

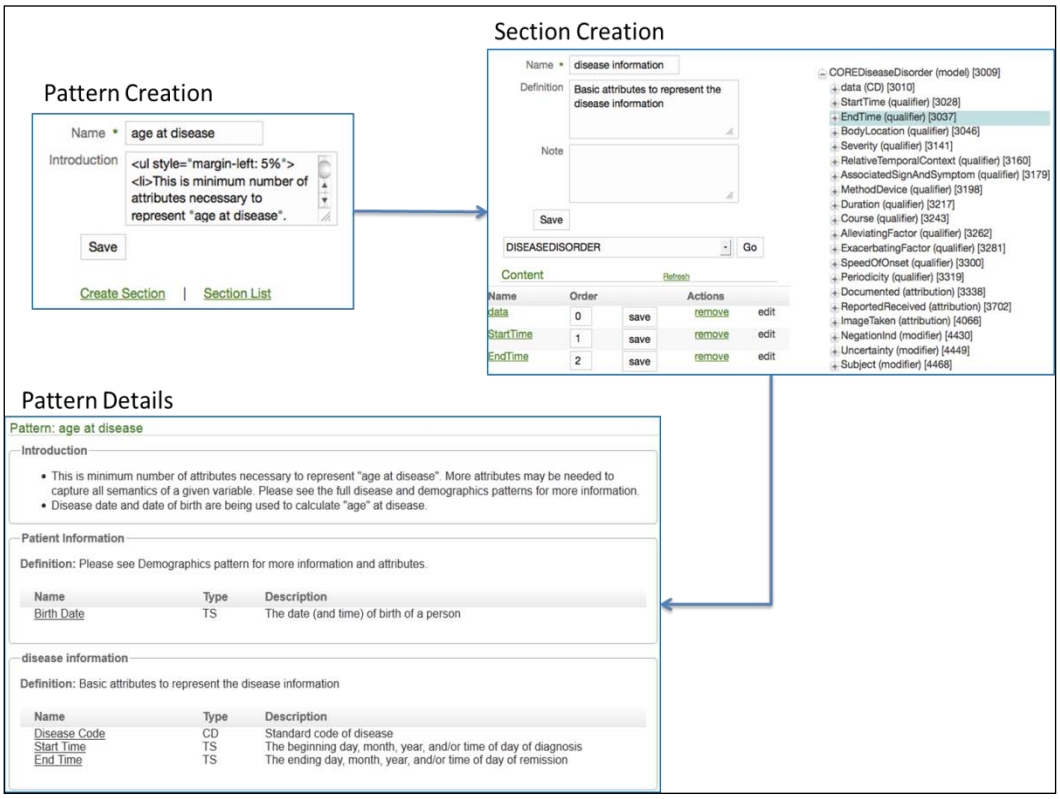


Figure 4: User interface for Pattern creation. The figure illustrates the workflow for creating a Pattern with the web-based user interface. The "age at disease" pattern is shown as an example.

Application and Evaluation

We evaluated this approach by extending our previous work on the semantic harmonization of data dictionaries from the Pharmacogenomics Research Network (PGRN)[7]. Due to the large size of the data set, RRF and QZ grouped PGRN data elements into arbitrarily-defined categories[2] to facilitate review and pattern identification. The data elements in each category were carefully reviewed and candidate patterns were identified, created, and documented using the process described above for each group of semantically similar data elements. This process was repeated until the majority of the PGRN data elements were mapped to a Pattern and a variety of Patterns were created.

The ability to use Patterns to capture the semantics represented by the PGRN data elements was evaluated by reviewing the mappings between each data element and its respective Pattern. In particular, QZ and RRF determined how each attribute would be used, and how they might be combined, to represent a given PGRN data element. The results were recorded as instructions for how the Pattern could be used to represent a given data element. If the initial definition of a Pattern could not fully represent a PGRN data element, the Pattern was extended as needed.

Results

The Pattern approach developed by this project was intended to provide an intuitive way for investigators to use standardized but highly complex CEMs for data collection without first acquiring specialized knowledge. To determine whether this approach could successfully represent the semantics of data elements used by research studies while still maintaining a link to the underlying formal models, which is necessary to enable transformation of the data into CEM syntax, we performed two evaluations. First, the mappings of 2,089 PGRN data elements that were previously mapped directly to CEMs were examined to verify that the semantics were retained when the data elements were mapped to one of the 16 Patterns created for this study (Table 1). All mappings were found to be semantically complete. Second, to demonstrate that the Pattern approach significantly reduced the complexity of the models by retaining only those attributes that were required to express the semantics of the data elements, the reduction in the number of attributes was quantified for each Pattern (Table 1). All Patterns reduced the number of attributes that an adopter would have to consider by >90%.

Pattern	Source CEMs	Number of Attributes		
		Source CEMs	Pattern	Difference (%)
Address	Patient	96	8	-88 (-92%)
Person Identifier	Patient	96	10	-86 (-90%)
Telecom	Patient	96	6	-90 (-94%)
Demographics	Patient, Primary Cause of Death	100	8	-92 (-92%)
Age at Disease	Disease/Disorder, Patient	221	4	-217 (-98%)
Disease	Disease/Disorder	125	12	-113 (-90%)
Disease History	Disease/Disorder	125	4	-121 (-97%)
Family History of Disease	Disease/Disorder, Personal Relationship Type	125*	6	-119 (-95%)
Drug Administration	Noted Drug	113	17	-96 (-85%)
Drug Admin. History	Noted Drug	113	5	-108 (-96%)
Laboratory Observation (Coded Result)	Lab Observation Coded	187	6	-181 (-97%)
Laboratory Observation (Quantitative Result)	Lab Observation Quantitative	190	6	-184 (-97%)
Blood Pressure	Systolic BP Meas., Diastolic BP Meas.	115	3	-112 (-97%)
Mean Arterial Pressure	Mean Arterial Pressure Meas.	115	2	-113 (-98%)
Heart Rate	Heart Rate Meas.	33	3	-30 (-91%)
Height Weight Measurement	Height Meas., Body Weight Meas., Body Mass Index Meas.	76	4	-72 (-95%)

Table 1: Patterns created for this study. The count of attributes from source models included the CEM elements of type item, modifier, qualifier, and attribution. Meas = Measurement. *The Personal Relationship Type model was not finished at the time of writing and therefore was excluded from the count.

Discussion and Conclusion

While CEMs provide a robust framework for the semantic representation and exchange of clinical data, their complexity can be a barrier to adoption. To facilitate the use of CEMs by investigators we sought to develop an approach that would present CEM content in a way that is more intuitive to non-informaticians while still retaining the full semantics of the underlying model. The Pattern approach permitted the creation of simplified, conceptual representations by hiding the complexity of the underlying CEMs that was not necessary for capturing the semantics of the data elements. Furthermore, the process of identifying generalized abstract data elements from groups of semantically similar, study-specific data elements resulted in the creation of relatively few Patterns, indicating that this approach is likely to scale well for all but the most highly-specialized data elements.

Current limitations of this approach include difficulty capturing highly qualified data elements (such as recording episodes of a disease that are associated with a specified condition) and modeling transformations that lead to derived values (e.g., logarithm), both of which are also limitations of the underlying CEM specification. These were addressed by using existing CEM attributes where possible (e.g., exacerbating factor) and by representing the pre-calculated value, respectively. Discussions with the model owners were required to establish consistent practices for representing complex concepts (e.g., pedigrees) and to extend the CEMs when needed, which may limit scalability. These issues could be mitigated by improved CEM documentation and tooling that supports community authoring.

To use CEMs directly, specialized knowledge is required to understand both the technical specification and how to interpret and implement the models themselves. The Pattern approach developed in this study eliminates those requirements for investigators by prespecifying, through hidden mappings to the underlying models, which attributes should be used to capture each aspect of a precoordinated, study-specific data element. The highly simplified models and intuitive instructions provided by Patterns lower the barrier for investigators to use standardized CEMs for data collection. The mappings between Pattern and CEM attributes permit the transformation of data into CEM syntax, which can then be used for data exchange or integration into an EMR. This approach may encourage the adoption of CEMs for data collection, thereby reducing the need for retrospective data normalization efforts.

Acknowledgment

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network). The authors thank Dr. Thomas Oniki for critical clarifications regarding the implementation of CEMs and Mr. Zonghui Lian for providing technical support.

*RRF and QZ contributed equally to this work.

References

1. Zhu Q, Freimuth RR, Lian Z, et al. Harmonization and semantic annotation of data dictionaries from the Pharmacogenomics Research Network: A case study. J Biomed Inform. 2013; 46(2):286-293.
2. Zhu Q, Freimuth RR, Pathak J, Chute CG. Using Clinical Element Models for pharmacogenomic study data standardization. Proc 2013 AMIA Summit on Clinical Research Informatics. 2013. 292-296.
3. Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. J Am Med Inform Assoc. 2007; 14(6):687-696.
4. Sansone S-A, Rocca-Serra P, Field D, et al. Toward interoperable bioscience data. Nat Genet. 2012; 44(2):121-126.
5. James A. Qualibria Constraint Definition Language (CDL) language guide. Oct. 22, 2010.
6. CEM Browser [Internet]. [cited 2013 October 10]. Available from: http://www.clinicalelement.com
7. Pharmacogenomics Research Network (PGRN) [Internet]. [cited 2013 October 10]. Available from: http://www.pgrn.org
8. Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform. 2012 Aug; 45(4):763-71.
9. Tao C, Jiang G, Oniki TA, et al. A semantic-web oriented representation of the clinical element model for secondary use of electronic health records data. J Am Med Inform Assoc. 2013; 20:554-562.

Ontology-Based Tools to Expedite Predictive Model Construction

Peter Haug, MD[1, 2]; John Holmen, PhD[1]; Xinzi Wu, PhD[1]; Kumar Mynam[1]; Matthew Ebert, MS[1]; Jeffrey Ferraro, PhD[1]

[1]Intermountain Healthcare, Salt Lake City, Utah; [2]University of Utah

Abstract

Large amounts of medical data are collected electronically during the course of caring for patients using modern medical information systems. This data presents an opportunity to develop clinically useful tools through data mining and observational research studies. However, the work necessary to make sense of this data and to integrate it into a research initiative can require substantial effort from medical experts as well as from experts in medical terminology, data extraction, and data analysis. This slows the process of medical research. To reduce the effort required for the construction of computable, diagnostic predictive models, we have developed a system that hybridizes a medical ontology with a large clinical data warehouse. Here we describe components of this system designed to automate the development of preliminary diagnostic models and to provide visual clues that can assist the researcher in planning for further analysis of the data behind these models.

Introduction and Background

Electronic medical record (EMR) systems have become a standard contributor to modern healthcare. In many clinical settings, they capture all of the information recorded as a part of the documentation of care. In the process, they provide tools to streamline medical processes as well as opportunities to enhance the medical decision making process both indirectly, through effective organization and presentation of patient data, and directly, through clinical decision support (CDS) technologies.

The clinical data that accumulates as a result of computer-assisted healthcare has an additional role to play. As it is collected and curated over time, it can be analyzed to yield insights into the diagnosis and treatment of disease. The process of mining this data can result in new medical knowledge that can lead to changes in care. It can also support new CDS-based interventions built upon models derived from a combination of clinical data and medical knowledge.

We believe that medical data mining focused on the development of clinically useful models has large potential value as a way to increase the ability of EMR systems to standardize and expedite care. To support research in this area, we have constructed a system whose focus is the use of data from a large enterprise data warehouse (EDW) combined with medical knowledge stored in a disease-oriented ontology. This combination is used to automate the construction of computable diagnostic models. We call this system the Ontology-driven Diagnostic Modeling System (ODMS)[1]. Here we will describe features of this system designed specifically to support a researcher as she/he generates and evaluates tools for real-time clinical diagnosis.

The system described below takes advantage of an extensive EDW created by Intermountain Healthcare, a large healthcare delivery system in Utah. This data warehouse captures data representing ~3 million visits each year collected while serving ~1 million patients. The data represents outpatient, inpatient, and emergency services and are a common focus for a variety of clinical research activities.

In recent years, we have brought added attention to the development of decision aids in the Emergency Department, notably a diagnostic and therapeutic CDS system for community acquired pneumonia[2]. The resulting system currently provides daily care in a group of 4 Utah hospitals. It is a prototype for the systems the ODMS is designed to produce.

The ODMS provides a group of embedded tools whose goal is to reduce the effort necessary to build diagnostic models. Key components of this process are addressed by the system including:

- Cohort development: the identification of patient subgroups with and without the target condition.
- Feature selection/extraction: the selection and extraction of those data elements relevant to the diagnostic problem.
- Model specification: the choice of a predictive modeling approach for the required diagnostic model.
- Model evaluation: tools that provide an initial exploration of the quality of the diagnostic models developed.

The goal of this system is to provide a semi-automated tool to generate and return to the user a collection of outputs designed to provide a good starting point from which to continue the construction of a functioning diagnostic model. The results of this initial evaluation assist the user in determining whether the available data in the EDW is indeed adequate to produce a useable clinical tool. Below we describe the components mentioned above and illustrated some of the features of the ODMS.

Methods

The ODMS uses as input the EDW and a diagnostic ontology. It outputs four key products. These outputs include:

- A diagnostic predictive model produced by the system.
- A list of the features found to be relevant to the diagnostic problem addressed.
- An analysis of the quality of the predictive model.
- The raw data generated by the system and used to develop the diagnostic model.

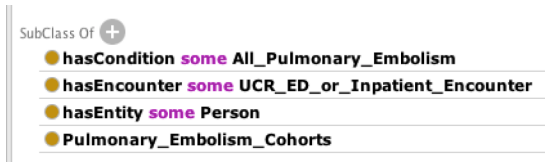


Figure 1: A piece of the disease ontology indicating the characteristics of patients who have PE. Displayed in Protégé.

These products are described and illustrated below. We will use a project where we have begun building a model for the diagnosis of pulmonary embolism (PE) to illustrate some of the features available in the ODMS.

A disease ontology is central to the operation of the ODMS and is used in the creation of the data sets necessary to the modeling effort. This ontology is designed to capture relationships that support identification of patient subgroups with and without the target disease and to identify clinical features important in the diagnosis of this condition. It has been developed largely through a combination of manual and semi-automatic introduction of taxonomies from various existing sources followed by the manual introduction of relevant properties connecting classes with the needed taxonomic components.

Figure 1 contains a fragment of the ontology defining a group of patients with PE. This fragment is displayed in the ontology management tool, Protégé[3]. It indicates the categories of patients for which to search. In this case, the ODMS uses taxonomic explosion to expand the "All_Pulmonary_Embolism" concept to include the 11 ICD-9 codes necessary to completely represent this concept. These concepts are harvested from the taxonomic trees embedded in the ontology (figure 2).

The ODMS also links to information that can tell the system which clinical environments to use when extracting a study cohort (in this case, ED or inpatient locations within a regional collection of care environments called the "Urban Central Region" (UCR)). In this case, it specifies the generic "Person" as the subject type, although further restrictions using age and sex are available.

The concepts in the ontology are linked, wherever possible, to standard national or international coding systems. For instance, diagnoses are linked to ICD-9 codes, laboratory results are linked to LOINC, and medications are links to RXNORM. Thus, when the ODMS reads through the ontology to construct a query against the EDW, the resulting queries are couched in terminologies that are largely standards-based. We anticipate that this will help us generalize this system to function in other settings.

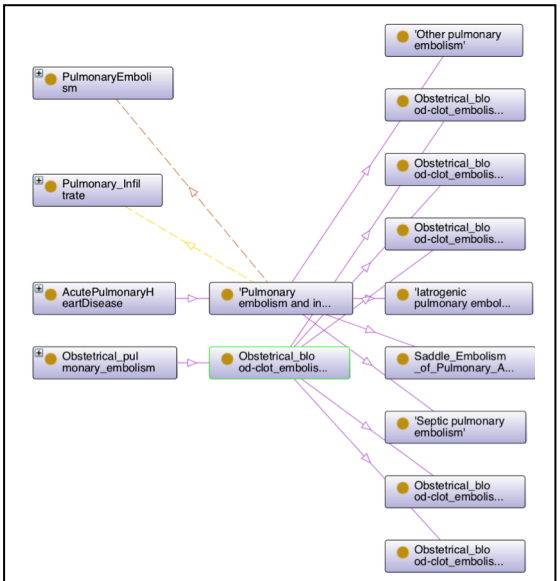


Figure 2: A portion of the diagnostic taxonomy embedded in the disease ontology used as a part of the ODMS. The ICD codes associated with pulmonary embolism are linked to the nodes in this tree structure.

Diagnostic Model

The ODMS is constructed to support the incorporation of a variety of predictive modeling tools within its framework. At present, we have incorporated four such tools. They are Naïve Bayesian Models, Tree-Augmented

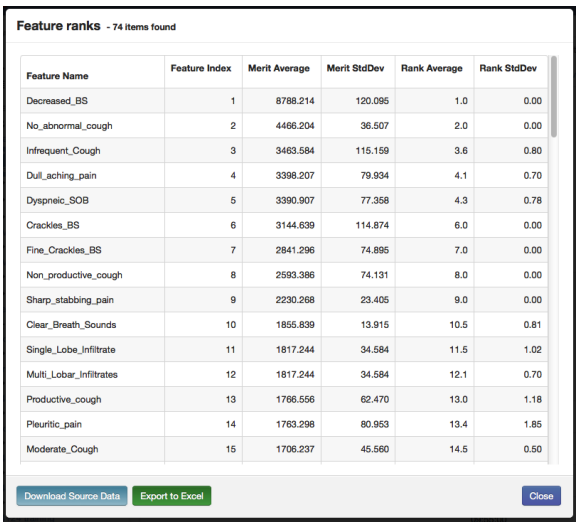
Bayesian Networks[4], K2-structured Bayesian Networks[5], and Random Forest Classifiers[6]. When appropriate, the system automatically discretizes continuous data using a Minimum Description Length algorithm (MDL)[7].

The modeling framework within the ODMS is architected to accommodate models from a variety of sources. Those listed above are based on components from Weka[8], a general purpose, data-mining environment; Netica[9], a Bayesian network tool; and custom predictive modeling tools that have resulted from local development efforts. When a new analysis is begun, the system provides a default protocol to guide the process. This can be reviewed and modified at system initiation. For users who do not wish to accept system defaults, alternate sets of components and procedures can be configured from a setup page.

Relevant Features

As indicated above, the ODMS uses ICD-9 codes extracted from an ontology-based, disease modeling system to identify groups of patients with and without the target diagnosis. The ontology also contains links from diseases to their diagnostic features. These include laboratory results, vital signs, x-ray results, nurse charting information, and chief complaints. The system interrogates the ontology for these relationships. They are used to build queries against data in the EDW and to retrieve this information for patients with and without the target diagnosis. As a part of this process, it assembles an exhaustive list of features, which typically includes dozens of variables that may contribute to the diagnostic model. In case of pulmonary embolism, this process generated seventy-four proposed variables to be used in the initial diagnostic model. Once the proposed data set is assembled, the ODMS activates a feature selection process designed to identify a subset of the features that will be the focus of further modeling efforts.

The ODMS uses an initial filtering step to reduce the number of features prior to model construction. The default number of variables to include is 15, but this parameter can be changed on the system configuration page. A simple testing procedure is used to rank these features by discrimination power. The system employs a Chi-squared algorithm to give an initial assessment of the degree of association between each feature and the disease. Figure 3 displays the ranked list used by the system to display these associations while figure 4 depicts the graphical output of this ranking procedure.



Feature ranks - 74 items found

Feature Name	Feature Index	Merit Average	Merit StdDev	Rank Average	Rank StdDev
Decreased_BS	1	8788.214	120.095	1.0	0.00
No_abnormal_cough	2	4466.204	36.507	2.0	0.00
Infrequent_Cough	3	3463.584	115.159	3.6	0.80
Dull_aching_pain	4	3398.207	79.934	4.1	0.70
Dyspneic_SOB	5	3390.907	77.358	4.3	0.78
Crackles_BS	6	3144.639	114.874	6.0	0.00
Fine_Crackles_BS	7	2841.296	74.895	7.0	0.00
Non_productive_cough	8	2593.386	74.131	8.0	0.00
Sharp_stabbing_pain	9	2230.268	23.405	9.0	0.00
Clear_Breath_Sounds	10	1855.839	13.915	10.5	0.81
Single_Lobe_Infiltrate	11	1817.244	34.584	11.5	1.02
Multi_Lobar_Infiltrates	12	1817.244	34.584	12.1	0.70
Productive_cough	13	1766.556	62.470	13.0	1.18
Pleuritic_pain	14	1763.298	80.953	13.4	1.85
Moderate_Cough	15	1706.237	45.560	14.5	0.50

Download Source Data | Export to Excel | Close

Figure 3: A ranked list of the diagnostic features to be used in development of the diagnostic model. The features in the list are those that will be used to suggest the diagnosis of pulmonary embolism prior to definitive testing.

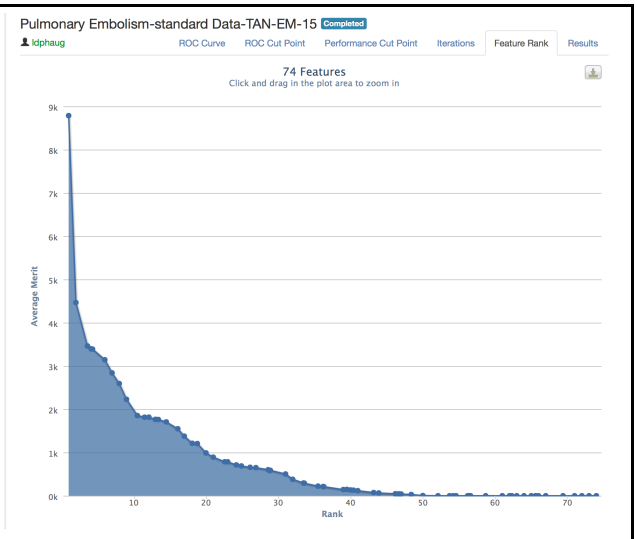


Figure 4: A graphical representation of the ranking for features based on their discriminating power (Chi-square score) for diagnosing pulmonary embolism. Within the 74 extracted features the Chi^2 approaches 0 after the 40^{th} variable.

The Chi-squared test statistic is used as an initial filter for the proposed features. The user determines the number of variables that should be included in the analysis from this feature selection process. Additional feature selection algorithms may further reduce the number of variables as a part of the model generation process.

Model Evaluation

As a part of its initial analysis, the ODMS produces the predictive model specified at the start of the procedure. It also does an initial evaluation of this model. This is useful to reassure the user of the validity of the results. Issues

such as over-fitting are typically avoided in this way. Two algorithms are offered as a part of this initial evaluation step, N-fold cross validation and bootstrapping. In each case the system defaults to a 10-fold approach, but the user can modify this to include any number of steps deemed appropriate.

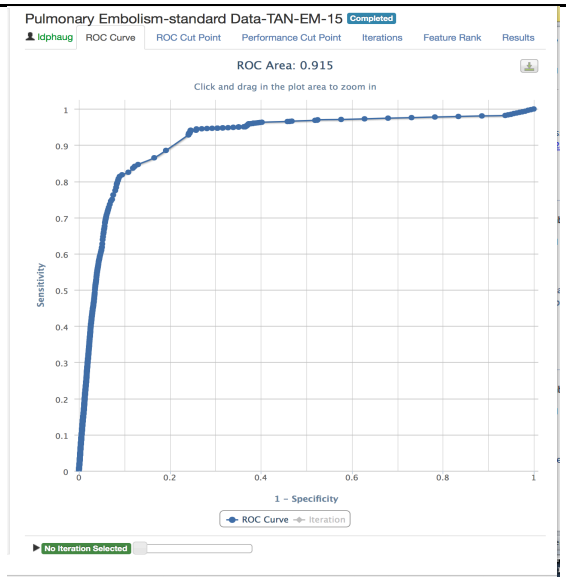


Figure 5: The ROC curve for a pulmonary embolism, predictive diagnostic system developed using emergency department data. This graph displays the result of an initial run of the ODMS.

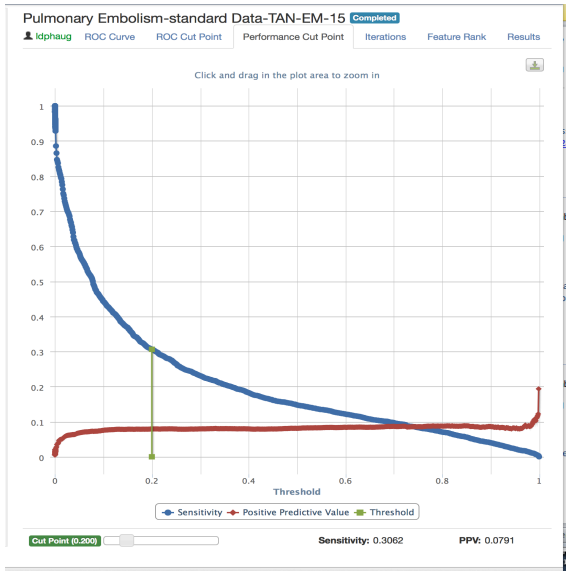


Figure 6: This graph compares recall (sensitivity) and precision (positive predictive value) over a range of thresholds for a Bayesian diagnostic system for pulmonary embolism.

Visualization is a vital part of model evaluation. For modeling tools that return a numeric value representing the likelihood that a case will fit a specific category (diagnosis), the system provides a number of graphical outputs including ROC curves (figure 5), recall vs. precision graphs (figure 6), and others. The goal is to both give an overall sense of the quality of the predictive model and to allow exploration of the characteristics of this model. Tools of this sort allow the user to estimate the operating characteristics that a diagnostic system will have when used in a real-world clinical setting.

The ODMS not only produces an initial predictive model with a minimum of user effort, it also supports refinement of this model under control of the user. To support this process of refinement, the ODMS includes graphical tools for visually comparing different models. Figure 7 compares two ROC curves from an initial and a refined diagnostic model developed for pulmonary embolism.

Raw Modeling Data

As mentioned above, the ODMS returns the raw data used in the initial modeling effort along with the model and evaluation. This allows the system's user to inspect this data, to modify it when appropriate (e.g. create derived variables, reduce redundancy, etc.) and to resubmit it to the modeling system for re-evaluation. The user may also process this data using other data mining systems that provide access to predictive algorithms not yet available in the ODMS.

The data extracted from the enterprise data warehouse is represented in the form of an ARFF file[10]. This standard format is compatible with a number of data mining tool kits, notably Weka[11], an extensible data-mining framework from New Zealand. These files are easily converted to other standard data analysis formats. We have found this capability to be useful for testing new components prior to adding them to the ODMS toolkit.

Results

The ODMS is capable of providing valuable assistance in automating data mining processes focused on the creation of diagnostic systems. We are currently using the system in modeling efforts for pneumonia, sepsis, and pulmonary embolism. The ODMS's major drawback is that our disease ontology is incomplete. Initial development efforts were directed toward demonstrating the value of this knowledge source as a tool in predictive data mining. The initial focus was in pulmonary diseases. We are now extending the ontology to encompass other diagnostic categories of interest to local research communities.

Figure 7: A comparison of two predictive algorithms for pulmonary embolism. The ROC curves depicted include one from an initial Bayesian network model created automatically and a second model produced after refinement of the data set extracted by the ODMS.

Discussion

The use of ontologies to expedite clinical research is attractive for several reasons. First, a great deal of fundamental medical knowledge can be encoded in ontologies. A variety of knowledge assemblies have been created that capture important taxonomic relationships. These relationships are readily available in terminologies like ICD-9, SNOMED, LOINC, etc. To extend these structured terminologies we have been working to introduce meaningful links between the concepts represented (disease leads to clinical observation, medication treats disease, etc.). Many of these links are available in some form already; bringing them together into a unified ontology is the logical next step.

But an ontology that captures the relationships relevant to clinical research is insufficient. The key to model building is to link the concepts in the ontology to data in a clinical data warehouse. These links allow the ontology to direct the collection of data relevant to a particular diagnostic modeling problem.

In future versions of the system described here, we hope to both extend the ontology and to improve linkage to our enterprise data warehouse. Toward this end, we have been constructing a special collection of data derived from our EDW. We refer to this as the Analytic Health Repository and it has the expressed goal of tying all data possible to national or international medical terminologies. These are the same terminologies used in the disease ontology, and we anticipate that this approach will make the ODMS easier to enhance and maintain. We also hope that integrating knowledge stored in ontologies with the clinical experience represented by large data warehouses will allow us to provide medical researchers with an efficient and effective environment in which to ask important medical questions.

Supported in part by Grant LM010482 from the National Library of Medicine.

References

1. Haug PJ, Ferraro JP, Holmen J, et al. An Ontology-Driven, Diagnostic Modeling System. J Am Med Inform Assoc Published Online First: [March 24, 2013] doi:10.1136/amiajnl-2012-001376.
2. Dean NC, Jones BE, Ferraro JP, et al. Performance and utilization of an emergency department electronic screening tool for pneumonia. JAMA Intern Med. Published Online first: 18 Mar 2013.
3. Protégé-OWL: http://protege.stanford.edu/ overview/protege-owl.html (accessed 21 Dec 2012).
4. Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. Machine Learning 1997;29:131–63.
5. Cooper GF, Herskovits E. A Bayesian Method for the Induction of Probabilistic Networks from Data. Machine Learning, 9, 309-347 (1992).
6. Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.
7. Fayyad UM, Irani KB. Multi-interval discretization of continuous valued attributes for classification learning. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93); Chamberry, France. 1993:1022–7.
8. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 2009;11:10–8.
9. Netica Software. Norsys Software Corporation; [March 16, 2011]. http://www.norsys.com (accessed 21 Dec 2012).
10. ARFF:http://weka.wikispaces.com/ARFF (accessed 21 Dec 2012).
11. Weka:http://www.cs.waikato.ac.nz/ml/weka/ (accessed 21 Dec 2012).

Extracting and standardizing medication information in clinical text – the MedEx-UIMA system

Min Jiang, MS[1], Yonghui Wu, PhD[1], Anushi Shah, MS[2], Priyanka Priyanka, BAMS[3], Joshua C. Denny, MD, MS[2], Hua Xu, PhD[1]
[1]School of Biomedical Informatics and [3]School of Public Health, The University of Texas Health Science Center at Houston, TX, US
[2]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, TN, US

ABSTRACT
Extraction of medication information embedded in clinical text is important for research using electronic health records (EHRs). However, most of current medication information extraction systems identify drug and signature entities without mapping them to standard representation. In this study, we introduced the open source Java implementation of MedEx, an existing high-performance medication information extraction system, based on the Unstructured Information Management Architecture (UIMA) framework. In addition, we developed new encoding modules in the MedEx-UIMA system, which mapped an extracted drug name/dose/form to both generalized and specific RxNorm concepts and translated drug frequency information to ISO standard. We processed 826 documents by both systems and verified that MedEx-UIMA and MedEx (the Python version) performed similarly by comparing both results. Using two manually annotated test sets that contained 300 drug entries from medication list and 300 drug entries from narrative reports, the MedEx-UIMA system achieved F-measures of 98.5% and 97.5% respectively for encoding drug names to corresponding RxNorm generic drug ingredients, and F-measures of 85.4% and 88.1% respectively for mapping drug names/dose/form to the most specific RxNorm concepts. It also achieved an F-measure of 90.4% for normalizing frequency information to ISO standard. The open source MedEx-UIMA system is freely available online at http://code.google.com/p/medex-uima/.

INTRODUCTION

Electronic Health Records (EHRs) are becoming an enabling resource for drug outcome studies.1 However, medication data are often recorded in heterogeneous formats in EHRs. With the increased use of computerized provider order entry (CPOE) systems, electronic prescribing (e-prescribing) tools, and electronic medication administration record systems (e-MARs), medication records in the EHR are increasingly available as structured entries. However, much current and historical medication information is still embedded in narrative text entries within clinical documentation, patient problem lists, or communications with patients through telephone calls or patient portals, especially in the outpatient settings. Therefore, natural language processing (NLP) methods that can extract medication information from clinical narratives and encode them into standard representations have received great attention, as detailed below.

Early studies primarily focused on extracting drug names from clinical notes. In 1996, Evans et al. built the CLARIT2 system to extract the drug name and dosage phrases in discharge summaries and reported an accuracy of 80%. Chhieng et al.3 reported a precision of 83% by using a string matching method to identify drug names in clinical records. In 2009, Jagannathan et al.4 evaluated the performance of four commercial clinical NLP systems on medication information extraction (including drug names, strength, route, and frequency). These systems demonstrated high F-measures (93.2%) for capturing drug names, but lower F-measures (85.3%, 80.3%, and 48.3% respectively) on retrieving strength, route, and frequency. In 2009, Informatics for Integrating Biology and the Bedside (i2b2), an NIH-funded National Center for Biomedical Computing (NCBC) based at Partners Healthcare System in Boston, organized an clinical NLP challenge to extract medication names and their associated signature fields including dosage, mode, frequency, duration, and reason from hospital discharge summaries.5 Twenty teams from twenty-three organizations and nine countries participated in the challenge. A variety of medication information extraction systems were developed and included systems using rule-based,6 machine learning based,7,8 and hybrid approaches,9 with overall promising results.

Despite the active NLP work on medication extraction, most of existing systems output medication related entities as textual fields, without mapping to standard representations such as RxNorm[10] for drugs and ISO 8601 standard for frequency information. One study done by Levin and colleagues[11] developed an effective rule-based system to extract drug names from anesthesia records and map to RxNorm concept unique identifiers (RxCUIs), with 92.2% sensitivity and 95.7% specificity. However, this study focused on encoding drug ingredients/brands only. In the example "Cetirizine 5 mg oral tablet", Levin's system will only encode the drug name "Cetirizine" (RxCUI 20610). However, an RxNorm concept actually can include three components: drug name (generic or brand), dose, and form. For the above example, a more specific RxCUI (1014676 – "cetirizine hydrochloride 5 MG Oral Tablet") could be assigned. With available drug dose and form (and/or route) information extracted by NLP systems, more specific RxCUIs can be assigned to medications in clinical text, which can be useful for other

computerized applications. For example, the dose form (e.g., intravenous vs. oral vs. topical) can imply very different indications and side effects. Frequency information is also important for medications and different string variants can often represent the same frequency (e.g., "two times a day" is equivalent to "b.i.d"). Therefore, normalization of drug frequency information is needed. However, few clinical NLP systems provide normalized frequency values. In the 2012 i2b2 NLP challenge on temporal information extraction, temporal expressions including frequency were normalized based on the ISO 8601 standard as in the TIMEX3[12] tag, which is the part of TimeML, a formal specification language for events and temporal expressions. To the best of our knowledge, TIMEX3 normalization has not been applied to the extraction of drug frequency information in clinical text.

In previous work, we developed MedEx,[13] a Python-based NLP system which could extract drug names and signature information with over 90% F-measure in discharge summaries and clinical visit notes from Vanderbilt University Hospital. We applied an extended version of MedEx to the 2009 i2b2 NLP challenge on medication extraction; it was ranked as the second best system among twenty entries.[6] We also developed simple normalization modules for dose and frequency, and integrated them with MedEx to calculate daily dose of tacrolimus[14] and weekly dose of warfarin.[15] In this study, we re-implemented MedEx in Java, based on the Unstructured Information Management Architecture (UIMA),[16] which is a component software architecture for development, discovery, composition, and deployment of multi-modal analytics for unstructured data. We name the new system "MedEx-UIMA" and it is freely available as open-source software. We also developed two new components in MedEx-UIMA and evaluated them herein: 1) encoding drugs to specific RxNorm concepts and 2) normalizing frequency to TIMEX3 format.

METHODS

As shown in Figure 1, the MedEx-UIMA system consists of two main components: 1) an information extraction module, which extracts medication related fields from clinical text; and 2) a standardization module that encodes drug name/dose/form information into RxCUIs and normalizes frequency information to the TIMEX3 format. The information extraction module basically is a Java implementation of the previous Python version of MedEx, with additional changes in transformation and disambiguation. The RxCUI encoding and frequency normalization are new functionalities of MedEx-UIMA. They are the primary focus of this study.

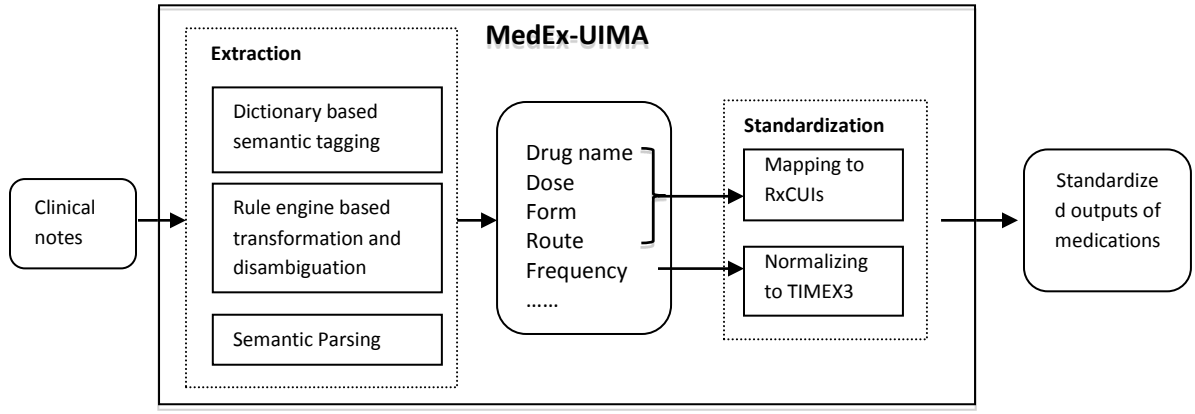


Figure 1. An overview of the MedEx-UIMA system

The UIMA implementation of MedEx

Using on the UIMA framework, we re-built the MedEx in Java as a pipeline-based system, where we defined classes including Sentence Boundary Detector, Tokenizer, Section Tagger, Semantic Tagger, Parser and Encoder. One significant change to the new MedEx-UIMA system is that we applied the Drools rule engine (http://www.jboss.org/drools/) to handle heuristic rules used in semantic tagging for tag transformation and word sense disambiguation. The rule-engine separates rule management from the workflow, thus making it possible for non-technical users to modify rules needed for specific tasks. The encoder is a new component in MedEx-UIMA, which maps drug name, dose, and form information to most specific RxNorm concepts and normalizes frequency information to TIMEX3 format.

Mapping drug name, dose, and form information to RxNorm concepts

When encoding drug information extracted from clinical text using RxNorm, there are two primarily options: 1) least-specific: map drug names only, e.g., to generic names such as "cetirizine"; or 2) most-specific: map to more specific RxNorm concepts that could contain drug name (either generic or brand), dose, and form information, such as "Cetirizine 5 MG Oral Tablet." In MedEx-UIMA, we provide both types of RxCUIs for a given drug entity. It is straightforward to map drug names to least specific

RxCUIs (the generic ingredient). We created a mapping between brand names and generic names based on RxNorm relationships and built a simple dictionary lookup function to map extracted drug names to their corresponding generic name RxCUIs.

Determining the most specific RxCUIs based on extracted drug name, dose, route, and form information is more challenging. We developed a rule-based approach for this task, which consists of four steps:

1. *Normalize drug information extracted by MedEx*: Five fields extracted by MedEx including drug name, dose, dose amount, route, and form are used to generate normalized fields of drug name, dose, and form. The normalization process is based on heuristic rules and manually created knowledge bases. In the example of "Cetirizine 5000 mcg tabs", MedEx will recognize "cetirizine" as a drug name, "5000 mcg" as the dose, and "tablet" as the form. The normalization program will produce normalized results as (Generic name: cetirizine), (Dose: 5 mg), and (Form: tablet), which can then be mapped to the RxNorm entry. In this example, rules for conversion between different units in the dose field and knowledge for recognizing "Tab" and "Tablet" as synonyms were used in the normalization process. We have developed knowledge bases about synonyms and route-form mappings for normalizing drug forms.
2. *Normalize drug information of RxNorm concepts*: For each RxNorm term, we process it using the same procedure as in step 1 and generate normalized fields for drug names, dose, form etc.
3. *Generate RxNorm candidate entries*: For a given drug entry, we search all RxNorm concepts and generate a list of candidate concepts containing the same normalized drug name.
4. *Rank RxNorm candidate concepts by calculating similarity scores between the normalized drug entry and candidate concepts*: Once the drug name, dose and form information is normalized, we concatenate them in an order to generate a string. We then calculate weighted Jaccard Similarity17 scores between a drug entry string and all its corresponding candidate string. The Jaccard Similarity is defined as the ratio between the number of common words in both two strings, multiplied with the weight of each word, and the number of words in any of two strings ,multiplied with their weight. We assign different weights to different drug fields to reflect their search priorities. For example, the default weight of any word is "1". But we assign a higher weight (e.g., 1.8) to the dose field, as the same dose is a strong indicator. The RxNORM concept with the highest similarity score with the drug entry is then selected as the most specific RxNORM code.

Figure 2 shows an example of searching the most specific RxNORM codes. As shown in the figure, "Augmentin 200-28.5 MG Oral Tablet" is the input sentence. Drug (Augmentin), dose (200-28.5 MG) and form (Oral Table) are extracted and normalized by MedEx. After searching the drug name "Augmentin", multiple RxNORM candidate entries are generated, including "Augmentin, 200 mg-28.5 mg oral tablet, chewable", "Augmentin, 200 mg-28.5 mg/5 mL oral powder" etc. All RxNORM candidate entries are normalized in the same way. Then we calculate Jaccard similarity between the string "Augmentin 200-28.5 MG Oral Tablet" and each of the RxNORM candidate strings. The one with highest similarity score is then selected as the most specific RxNORM entry.

Normalizing drug frequency information to the TIMEX3 format

The frequency normalization module was constructed on our temporal expression extraction system developed for the 2012 i2b2 NLP challenge on temporal information extraction. The original system is a rule-based system developed in Python to extract three types of temporal expressions, including date, frequency and duration. We re-implemented the system in Java and extended it with new regular expression rules for handling additional drug frequency patterns observed in the development set. The following example shows how the rule-based system normalizes frequencies into TIMEX3 format. For the expression "three times per week", the frequency normalization module first detects the normalizable components using the rule "(%NumWord) (%TIMES) (%PER) (%DayUnit)". Strings starting with "%" are predefined patterns using regular expressions, where "NumWord" is a lexicon of all the possible numbers in English words, "TIMES" is a lexicon of all the possible expression for times (e.g., "times", "x"), "PER" is a lexicon of all the possible expression for every (e.g., "every", "per", "each"), and "DayUnit" is a lexicon of all the possible units of days (e.g., "day", "week", "month"). Once the regular expression is triggered, the normalization rules will be applied to normalize the NumWord "three" into "3", DayUnit "week" into "W" to generate the normalized value "R3P1W", where R stands for "Repeat" and P stands for "Period." One difference between our drug frequency normalization and the i2b2 challenge guidelines was that we do not average a range (e.g., the i2b2 guidelines normalize "three to four weeks" into "P3.5W"; however, our system normalizes it as "P3-4W")

Evaluation

We first compared the performance of the MedEx-UIMA with the previous Python-based MedEx system (MedEx-Python). We processed 826 clinical notes from the 2010 i2b2 challenge using both MedEx-Python and the MedEx-UIMA systems. We then took the outputs of MedEx-Python as the gold standard and calculated precision/recall/F-measure of MedEx-UIMA against the gold standard. In addition, we reviewed 100 randomly selected discrepant drug entities by the two systems and counted the number of correct samples by MedEx-UIMA.

To develop and evaluate the encoding modules for drug name and frequency information, we created manually annotated datasets. We first used the dataset from the 2009 i2b2 clinical NLP challenge, which was to extract medication information from discharge summaries. The i2b2 dataset contains 251 discharge summaries collectively annotated by challenge participants, in which drug names and associated strength, route and frequency information were identified. We randomly divided the dataset

into two subsets: 126 notes as the development set and 125 notes as the test set. From the development set, we collected all i2b2 annotated drug entities and annotated 300 randomly-selected distinct drug entities. These 300 drug entities (with their sentences) were used to develop our system.

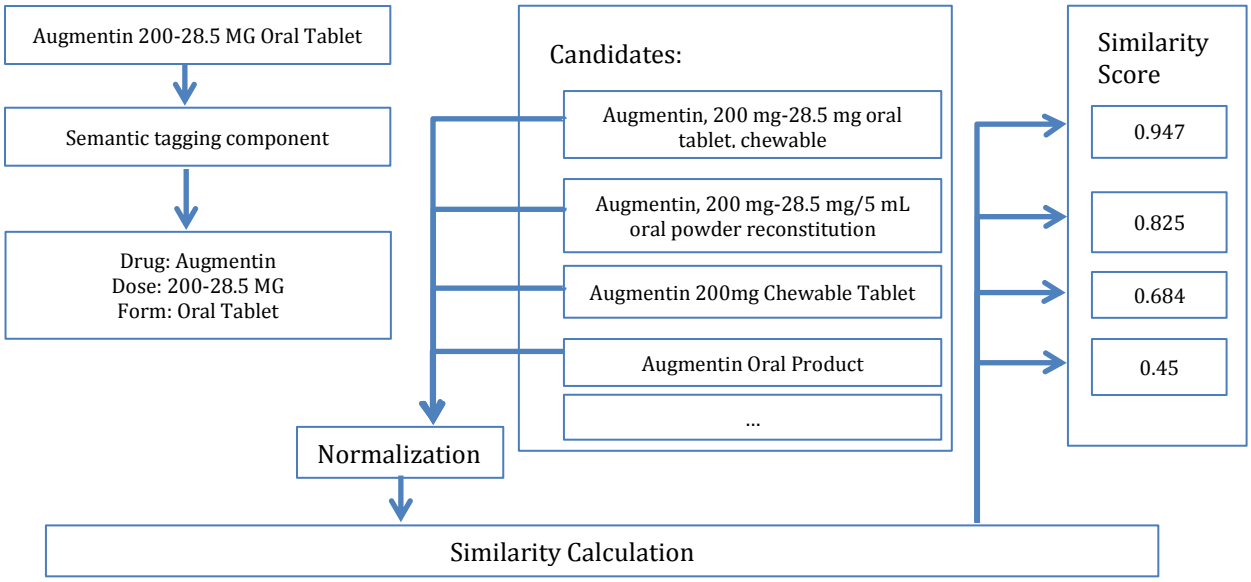


Figure 2. The example of determination of most specific RxNORM code

From the test set, we also collected all drug entities and randomly selected 300 drugs for annotation, which served as the independent test set to evaluate our system. For each drug entity in the development and test set, the original sentence containing the drug as well as drug name, dose, and route fields extracted by the i2b2 challenge, were presented to a medical domain expert for manual review. To encode RxNorm concepts, the annotator searched RxCUIs using RxNav, which is graphical search interface for RxNorm concepts. For frequency normalization, the annotator manually entered the normalized value for each frequency expression. In addition to the i2b2 dataset, which primarily contains drug entries in clinical narratives, we generated another test set containing more structured medication data. We randomly selected a list of 300 medications entries from computerized order entry system at UT Physician, a clinic of University of Texas Health Science Center at Houston, and manually annotated them with RxNORM codes following the same procedure.

We evaluated the performance of our system by reporting standard precision, recall, and F-measure on the independent test sets. For the first dataset, as the i2b2 challenge included drug classes such as "antibiotics", not all 300 drug entities in the test set can be coded by RxNorm concepts. Based on manual review, 270 drugs in the test set were classified as codable drugs. Among 270 codable drugs, true positives were defined as samples that were extracted by MedEx-UIMA and assigned correct RxCUIs. Recall was defined as the ratio between the number of true positives and the total number of codable drugs (270). Precision was defined as the ratio between the number of true positives and the number of codable drugs recognized by MedEx-UIMA. Similar definitions were used to measure precision and recall for frequency normalization as well. There were 243 frequency expressions in the independent test set. To be qualified as true positives, a frequency expression must be recognized by MedEx-UIMA and assigned the correct normalized values in the TIMEX3 format. For the medication list from UT Physician, all three hundred medication entries were codable.

RESULTS

When the outputs of MedEx-Python served as gold standard, the MedEx-UIMA achieved a precision of 95.8%, a recall of 98.0%, and an F-measure of 96.9%, for recognizing all drug related fields including name, dose, route, frequency etc. Manual review of 100 discrepant results by two systems showed that 42% were judged better in MedEx (Python), and 58% were judged better in MedEx-UIMA. Thus, overall we estimate that MedEx-UIMA slightly outperforms the original version of MedEx in precision.

Table 1 shows the performance of MedEx-UIMA on extracting and encoding medication information using the independent test set. For mapping drug names to generic ingredients (least specific RxCUIs), on both the medication list and clinical narratives,

the system achieved F-measure (98.5% and 97.5% respectively), which was consistent with previously reported high performance of MedEx on recognizing drug names. Mapping to the most-specific RxCUIs (taking dose and form into consideration) was more challenging: MedEx-UIMA achieved a precision of 85.8% and recall of 85.0% on drugs from medication list and 89.3% and 87.0% on drugs from clinical narratives. For frequency normalization, our system reached a high F-measure of 90.4% (precision 91.9% and recall 88.9%) on clinical narratives.

Table 1. Evaluation results of MedEx-UIMA on extracting and encoding drug and frequency information

Tasks	Precision	Recall	F-measure
Drug encoding - least-specific RxCUIs (Clinical narratives)	98.8%	96.3%	97.5%
Drug encoding - most-specific RxCUIs (Clinical narratives)	89.3%	87.0%	88.1%
Frequency normalization (Clinical narratives)	91.9%	88.9%	90.4%
Drug encoding - least-specific RxCUIs (Medication list)	99.0%	98.0%	98.5%
Drug encoding - most-specific RxCUIs (Medication list)	85.8%	85.0%	85.4%

DISCUSSION

In this study, we re-implemented MedEx, a high performance medication information extraction system, in Java using the UIMA framework. Evaluation showed the MedEx-UIMA system had similar high performance on recognizing drug related entities as MedEx (Python version). We also extended the encoding function of MedEx-UIMA to map drug names to generic ingredients and also the most specific RxNorm concepts, and developed a module to normalize frequency expressions to the standard TIMEX3 format. Our evaluation using a test set from the 2009 i2b2 challenge demonstrated that MedEx-UIMA can extract and encode drug name and frequency information with good performance. Such standard medication information extracted from clinical text can not only facilitate EHR-based clinical and translational research, but can also benefit computerized clinical applications such as clinical decision support systems and medication reconciliation processes. More importantly, MedEx-UIMA is available to the public as an open-source system, which can be freely downloaded from Google Code at http://code.google.com/p/medex-uima/.

We analyzed errors in mapping drug name/dose/form to RxNorm Concepts. Recall errors were often caused by unrecognized synonyms, abbreviations, or misspelled words. For example, “MVI” is a common abbreviation for “Multi-Vitamins”; but it could not be mapped to the expanded name by MedEx-UIMA, thus no RxCUI could be assigned. Precision errors had two primary causes. One is related to insufficient rules or knowledge for normalizing drug name, dose, and form information extracted by the NLP system. For example, “regular insulin” was not mapped because we did not add the fact of “regular insulin” = “insulin” to our knowledge base. The other regards selecting the correct RxCUI from multiple candidate concepts. Our current approach relies on simple string matching between drug name, dose, and form fields. More sophisticated code selection methods will be investigated in future development. For example, we plan to look into information retrieval methods to rank candidate concepts based on the querying drug string.

This study has limitations. One of them is the annotation process, which only involved one annotator, with some oversight and review of unclear cases by a board-certified internist. We plan to recruit multiple annotators for future development so that we can reduce bias introduced by annotation. The evaluation of drug name encoding and frequency normalization was based on selected drug entities at sentence level. In the future, we plan to further evaluate the performance of MedEx-UIMA at the clinical document level. Another limitation is that only documents from the i2b2 challenge were used; future studies should examine more documents types from other institutions.

CONCLUSION

In this study, we developed MedEx-UIMA, an open source medication information extracting and encoding system based on the existing MedEx system. It not only recognizes medication related entities with high performance, but also encodes drug names to specific RxNorm concepts and frequency information to ISO standard. Such a tool will have broad uses in various clinical settings, as well as EHR-based clinical and translational research.

ACKNOWLEDGEMENT

This study was supported in part by National Institute of General Medical Sciences grant 1R01GM102282, National Cancer Institute grant R01CA141307, Cancer Prevention & Research Institute of Texas grant RX1307, and the Office of the National Coordinator for Health Information Technology grant No. 10510592 for Patient-Centered Cognitive Support under the Strategic

Health IT Advanced Research Projects Program (SHARP). We would like to thank organizers of the i2b2 clinical NLP challenges for providing the annotated data sets for research uses.

References

1. Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. *Clinical pharmacology and therapeutics.* Mar 2011;89(3):379-386.
2. Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc AMIA Annu Fall Symp.* 1996:388-392.
3. Chhieng D, Day T, G G. Use of natural language programming to extract medication from unstructured electronic medical records. *AMIA Annu Symp Proc.* 2007;908.
4. Jagannathan V, Mullett CJ, Arbogast JG, et al. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int J Med Inform.* Apr 2009;78(4):284-291.
5. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc.* Sep-Oct 2010;17(5):514-518.
6. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc.* Sep-Oct 2010;17(5):528-531.
7. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc.* Sep-Oct 2010;17(5):524-527.
8. Li Z, Liu F, Antieau L, Cao Y, Yu H. Lancet: a high precision medication event extraction system for clinical text. *J Am Med Inform Assoc.* Sep-Oct 2010;17(5):563-567.
9. Tikk D, Solt I. Improving textual medication extraction using combined conditional random fields and rule-based systems. *J Am Med Inform Assoc.* Sep-Oct 2010;17(5):540-544.
10. Medicine NLo. RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm/. 2009.
11. Levin MA, Krol M, Doshi AM, Reich DL. Extraction and mapping of drug names from free text to a standardized nomenclature. *AMIA Annu Symp Proc.* 2007:438-442.
12. Pustejovsky J, Castaño J, Ingria R, et al. TimeML: Robust Specification of Event and Temporal Expressions in Text. *Fifth International Workshop on Computational Semantics.* 2003.
13. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc.* Jan-Feb 2010;17(1):19-24.
14. Birdwell KA, Grady B, Choi L, et al. The use of a DNA biobank linked to electronic medical records to characterize pharmacogenomic predictors of tacrolimus dose requirement in kidney transplant recipients. *Pharmacogenet Genomics.* Jan 2012;22(1):32-42.
15. Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *J Am Med Inform Assoc.* Jul-Aug 2011;18(4):387-391.
16. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Nat. Lang. Eng.* 2004;10(3-4):327-348.
17 Jaccard P. The distribution of the flora in the alpine zone. New Phytologist 11(2):37-50

Discovering Associations Among Diagnosis Groups Using Topic Modeling

Ding Cheng Li, Terry Thermeau, Christopher Chute, Hongfang Liu
Mayo Clinic, Rochester, MN 55901, USA

ABSTRACT

With the rapid growth of electronic medical records (EMR), there is an increasing need of automatically extract patterns or rules from EMR data with machine learning and data mining technqiues. In this work, we applied unsupervised statistical model, latent Dirichlet allocations (LDA), to cluster patient diagnoics groups from Rochester Epidemiology Projects (REP). The initial results show that LDA holds the potential for broad application in epidemiogloy as well as other biomedical studies due to its unsupervised nature and great interpretive power.

Introduction

With the rapid growth of electronic medical records (EMRs), it becomes more and more essential to develop methods to automatically mine information from EMRs with machine learning and data mining techniques in a timely and accurate manner [1, 2].

Recently, Latent Dirichlet Allocations (LDA) [3] has gained popularity in diverse fields due to the fact that it holds great promise as a means of gleaning actionable insight from the text or image datasets. In natural langauge processing (NLP), LDA clusters both words and documents into topics by approximating word or term distributions [4]. As an unsuperivsed statistical model, LDA makes use of Bayseisan inference to update the probability estimates for a hypothesis.

As LDA does not require a priori knowledge but can generate good interpretative models, enjoy good portability [5] and meanwhile it has the flexibility of adding implicit as well as explict priors to build diverse models [6-9], it thus holds the potential for broad applications, such as comorbidity studies, drug repurposing, biological connections among diseases and so on in biomedical research [10]. In this paper, we propose to use LDA to identify associations among diagnosis code groups utilizing an epidemiology cohort, Rochester Epidemiology Projects (REP) [11], and aim to understand the comorbidities. The paper starts with the introduction of background and related work in section 2; it then presents experimental methods in section 3 where the experiment data is introduced and adapted topic modeling for diagnosis group associations and topic analysis approaches are illustrated respectively. Section 4 presents the results and what can be found from those topics. Finally, in section 5, we discuss potential expansions, existing limitations and how we can make more improvements.

Background and Related Work

Disease classification and grouping in epidemiology studies

In epidemiology, the three *C*s (cause, contribute and correlate) in studying disease etiology proposed by Green [12] have long been the principle. However, diseases can be related biologically or phenotypically. There are different approaches to group diseases. The first approach defines disease groups by the symptoms of the affected organ. This kind of grouping derives from observational correlation between pathological analysis and clincal syndromes [13]. With the development of novel quantitaitive approaches to network analysis and the explosion of currently avaiable genomic, transcrptomic, proteomic and metabolomic data sets, biological systems based on network has been applied to disease classifications [14].

The most popular disease classification systems used is the International Classification of Disease (ICD) [15, 16], which classifies diseases systematically based on the analysis of the general health situation of population groups. It is used to monitor the incidence and prevalence of diseases and other health problem and has become the standard diagnosis tool for epidemiology.

However, ICD classification can be too finer granualarity for clinical practice since the number of ICD codes is too large and the distinctions among some codes are not clear. The large number of ICD-9-CM (the 9th version, Clinical Modification) codes also makes statistical analysis and reporting difficult and time-consuming. the Agency for Healthcare Research and Quality (AHRQ) introduces Clinical Classification Software [17] (CCS) to cluster patient

diagnosis and procedures into a manageable number of clinically meaningful categories. This way, 14,000 diagnosis codes are reduced to 279 groups.

Topic modeling in boimedical informatics Specifications

In biomedical informatics, probabilistic topic modeling has been applied to patients' notes to discover relevant clinical concepts and relations between patients [18]. Angues et al. [19] applied unsupervised LDA to primary clinical dialogues for visualizing shared content in communication. Wang et al. developed BioLDA [20] to find complex biological relationships in recent PubMed articles. Wu and Xu [21] made use of LDA to rank gene-drug relationships in biomedical literatures based on Kullback-Leibler (KL) distance between topics derived from LDA. Bisgin et al. [12, 22] mined FDA drug labels using topic modeling. Fifty-two unique topics, each containing a set of terms, were identified and then the probabilistic topic associations were used to measure the similarity between drugs. Bian et al. [23] utilized the topic features to categorize the collections tweets into latent topics and those topics are used as features to train SVM prediction models for mining adverse effects labels. Newman et al. [24] and Bundachus and Tresp [25] employed topic models to interpret MeSH terms. Chen et al. [26] proposed to use LDA to promote ranking diversity for genomics information retrieval and they claimed that topic distributions of retrieval passages can help identify aspects more accurately. Chen et al. [27] extended LDA by including background distribution to study microbial samples. Under their setting, each microbial sample is a document and each functional element is a word. They found that estimating the probabilistic topic model can uncover the configuration of functional groups. All of those studies have shown the potentiality of topic modeling.

Experimental Methods

In this study, our main goal is to investigate the effectiveness of topic modeling in discvering assocations among disease groups. We first generate topic distribution for selected medical records for certain population and then the connections among disease groups are analyzed.

Rochester epidemiology project (REP) and data inclusion

The Rochester Epidemiology Project (REP) is a collaboration between health care providers in southeaster Minnesota, which involves Olmsted Medical Center, Mayo Clinic, Rochester Family Medicine Clinic and other medical care providers in southeastern MN. The REP is a unique records-linkage research infrastructure that has existed since 1966. It includes the medical records of all persons who have ever lived in Olmsted County, Minnesota between January 1, 1966 and the present, and who have given permission for their medical information to be used for research. Those persons comprise more than 500,000 unique individuals and more 6 million person years of follow-up through 2010. Historically, the Olmsted County population is less racially diverse then the US as a whole [11, 28] and similar to the state of Minnesota and surrounding states [29]. The REP data we use has been processed and saved as a matrix with rows being the patient ID and columns the diagnosis code group defined by AHRQ. There are 256 diagnosis code groups in total in our data. As an initial study, we only select 4644 patients who are above 65 and paid 80 visits over the chosen set of years for this study.

Topic modeling

Topic modeling is originally a tool for text analysis. Now, we adapt it to the association analysis of diagnosis group. In text analysis, LDA represents a document as a mixture of fixed topics. Under the context of our data, LDA represents a collection of patients as a mixture of fixed topics. Each topic z has the weight $\boldsymbol{\theta}_{\boldsymbol{z}}^{\boldsymbol{p}}$ in a patient p and each topic is a distribution over a finite vocabulary of diagnosis code groups, and each code group c has a probability $\boldsymbol{\phi}$ in topic z. Placing symmetric Dirichlet priors on $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, with $\boldsymbol{\theta} \sim \boldsymbol{Dirichlet}(\boldsymbol{\alpha})$ and $\boldsymbol{\phi}^{\boldsymbol{z}} \sim \boldsymbol{Dirichlet}(\boldsymbol{\beta})$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are hyper-parameters to control the sparsity of distributions, the generative model is given by:

$$\boldsymbol{c_i}|\boldsymbol{z_i}, \boldsymbol{\phi}_{\boldsymbol{c_i}}^{\boldsymbol{z_i}} \sim \boldsymbol{Discrete}(\boldsymbol{\phi}^{\boldsymbol{z_i}}), \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{C} \qquad \boldsymbol{\phi}^{\boldsymbol{z}} \sim \boldsymbol{Dirichlet}(\boldsymbol{\beta}), \quad \boldsymbol{z} = \boldsymbol{1}, \dots, \boldsymbol{K}$$

$$\boldsymbol{z_i}|\boldsymbol{\theta}^{\boldsymbol{p_i}} \sim \boldsymbol{Discrete}(\boldsymbol{\theta}^{\boldsymbol{p_i}}), \quad \boldsymbol{i} = \boldsymbol{1}, \dots, \boldsymbol{C} \qquad \boldsymbol{\theta}^{\boldsymbol{p}} \sim \boldsymbol{Dirichlet}(\boldsymbol{\alpha}), \boldsymbol{p} = \boldsymbol{1}, \dots, \boldsymbol{P}$$

where K is the total number of topics, C is the total number of diagnosis code groups in the patient collection, and p_i and z_i are the passage and the topic of the ith code group c_i respectively. Each code group in the vocabulary $c_i \in V = [c_1, c_2, \dots, C_C]$ is assigned to each latent topic variable z_i. Given a topic $z_i = k$, the expected posterior

probability $\hat{\theta}^{p}$of topic mixings of a given patient p and the expected posterior probabilities $\hat{\phi}_{c_i}^{p_i}$ of code group c_i are calculated as below.

$$\hat{\phi}_{c_i}^{z_i} = \frac{n_{c_i k} + \beta}{\sum_{j=1}^{C} n_{c_j k} + C\beta} \qquad \hat{\theta}^{p} = \frac{n_{pk} + \alpha}{\sum_{j=1}^{K} n_{pj} + P\alpha}$$

where $n_{c_i k}$ is the count of c_i in topic k, and $n_{p,k}$ is the count of topic k in patient p.
In this study, we used the LDA approach to obtain the parameter ϕ for every diagnosis code group. The topics were extracted by using the R package *topicmodels,* which is based on Blei et al [3].

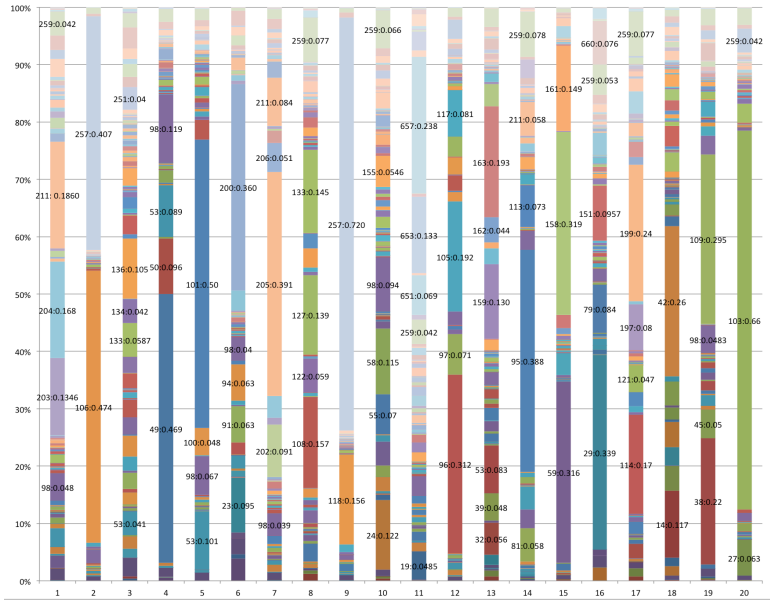


Figure 1 Diagnosis code group proportion for 20 topics where
x-axis is the topic and y-axis is the proportion of each code group in that topic

1.1. Associations discovery of diagnosis group

The topic distributions over diagnosis code group measures the connection (or relatednedss) of a disease with a specific topic (i.e. the conditional probability of topic for a given disease as shown in Figure 1. As shown in the previous section, in our work, the document is the patient while the term is the diagnosis code group. Therefore, the posterior distribution $\hat{\theta}$ would detemine the probability of a patentient given a topic and $\hat{\phi}$ would determine the probability of a diagonosis code group given a topic. More specifically, some patients were assigned to the most probable topics and some diagnosis code groups were assigned to the most probable topics.

Results and Analysis

There are a total of 4644 patients with their diagnosis code groups obtained with simeple exclusion criterias described above. LDA was employed to generate topic distributions for both the patients and the diagnosis code groups. We tested diverse topic numbers ranging from 20 to 147 and compared the resultant topics with respect to loglikelihood distributions and perplexities. Similar results were obtained when the number of topics is between 20 to 35. We chose the number of topics to be 20 and analyzed the common properties shared by the diseases with proportion higher than 0.05 in each topic.

Topic analasis in terms of disease relations

In Figure 1, the proportion of diagnosis code group for each of the topics is drawn with sample results when topic each topic is dominated by a few code groups which involve much larger ratios than remainings. Namely, each topic is represented by a few key diognosis code groups. In Table 1, the interpreations of those dominant diagnosis code groups are given. The five diagnosis code groups in T1 are almost relatedto joint disorders except the last two, *98* and *259*. T7 is also about joints, but it focuses more infections. The last two, are found in many topics. In fact, they two can be thought related to diverse diseases. That is why they have high proportions in many topics. Two components occupy 0.88 of T2 and T9. Both of them invovle the code *aftercare* while the other one for T2 is related to heart rhythm and the one for T9 is to infections in intravenes. It seems that these two topics are not clustered very well. But if we think from the perspective that *aftercare* plays imporant parts in quite a few severe diseases, especially diseases related to heart, it is quite reasonable for them two to co-occur often. T3 is obviously about respiratory diseases, with the four main codes nearly evenly distributed. Diseases in T4 seem to all related to fat-induced diseases since diabetes, hypertension, lipid metabolism may all be causes by eating too much high-

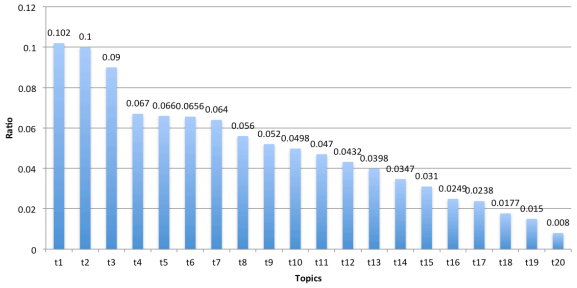


Figure 2 Patient ratio among topics

calory food. T8 all involves repiratory. Congestive heart failure and repiratory problems may be related. T5 and T12 are about heart diseases. number is 20. As can be seen, although each topic is composed of some proportion of all 253 diagnosis code group,

Nonetheless, T5 is more about heart organ itself while T12 is more about the circulation. T6, T11, T13, T14, T15, T17 and T18 have strong category features as sense, mental, nervous, urinary, system, kidney, skin and gastrointestinal diseases. T10 can be classified as internal secretion diseases. T16 seems more about dieseases seen among old people although data we used is in fact about patients who are older than 65. The results indicated that topic modeling can yield statistically significant topics that group and identify diseases sharing some commonality. Basically, what we have discovered about diseases, is consistent with what is shown by topic modeling for other domains, like text mining, natural langauge processing or image processing.

Topic analysis in terms of patient grouping

Figure 2 shows the distributions of patients' topic assignemtns. T1, T2 and T3 occupy about 0.3 among all topics and T4, T5, T6 and T7 also share about 0.06 respectivlye while T18, T19 and T20 only occupies about 0.017, 0.015 and 0.008 respectively. This is natural since the first seven diseases are all about heart diseases, respirative, tissue or joint disease which are quite common ones among old people. In contrast, the last three are about some rare diesease such as colon cancers, cerebrovascular or cancer of overy.

The actual counts of diagnosis code groups for each topic are somewhat different from the corresponding

Table 1 Corresponding diagnosis code group for each topic in **Figure 1**

Topic	AHRQ Clinical Classification Codes group and corresponding diseases																			
T1	211		204		203		98		259											
	Other connective tissue disease		Other non-traumatic joint disorders		Osteoarthritis		Essential hypertension		Residual codes; unclassified											
T2	106				257															
	Cardiac dysrhythmias				Other aftercare															
T3	136		133		134		53													
	Disorders of teeth and jaw		Other lower respiratory diseases		Other upper respiratory disease		Disorders of lipid metabolism													
T4	49		98		50		53													
	Diabetes mellitus without complication		Essential hypertension		Other endocrine disorders		Disorders of lipid metabolism													
T5	101		53		98		100													
	Coronary atherosclerosis and other heart disease		Disorders of lipid metabolism		Essential hypertension		Acute myocardial infarction													
T6	200		23		91		94		98											
	Other skin disorders		Other non-epithelial cancer of skin		Other eye disorders		Other ear and sense disorders		Essential hypertension											
T7	205		202		211		206													
	Spondylosis; intervertebral disc disorders; other back problems		Rheumatoid arthritis and related disease		Other connective tissue disease		osteoporosis													
T8	108		133		127		259		122											
	Congestive heart failure; no hypertensive		Other lower respiratory disease		Chronic obstructive pulmonary disease and bronchiectasis		Residual codes; unclassified		Pneumonia (except that caused by tuberculosis or sexually transmitted disease)											
T9	257				118															
	Other aftercare				Phlebitis; thrombophlebitis and thromboembolism															
T10	24		58		98		55		259		155		52							
	Cancer of breast		Other nutritional; endocrine; and metabolic disorders		Essential hypertension		Fluid and electrolyte disorders		Residual codes; unclassified		Other gastrointestinal disorders		Nutritional deficiencies							
T11	657		653		651		19		259											
	Mood disorders		Delirium, dementia, and amnestic and other cognitive disorders		Anxiety disorders		Cancer of bronchus; lung		Residual codes; unclassified											
T12	96		105		117		97													
	Heart valve disorders		Conduction disorders		Other circulatory disease		Peri-; endo-; and myocarditis; cardiomyopathy (except that caused by tuberculosis or sexually transmitted disease)													
T13	163		159		44		32		39		162									
	Genitourinary & ill-defined conditions		Urinary of urinary tract		Neoplasms of unspecified nature or uncertain behavior		Cancer of bladder		Leukemia		Other diseases of bladder and urethra									
T14	95		259		113		211		81											
	Other nervous system disorders		Residual codes; unclassified		Late effects of cerebrovascular disease		Other connective tissue disease		Other hereditary and degenerative nervous system conditions											
T15	158		59		161															
	Chronic kidney disease		Deficiency and other anemia		Other diseases of kidney and ureters															
T16	29		151		79		660		259											
	Cancer of prostate		Other liver disease		Parkinson's disease		Schizophrenia and other psychotic disorders		Residual codes; unclassified											
T17	199		114		197		259		121											
	Chronic ulcer of skin		Peripheral and visceral atherosclerosis		Skin and subcutaneous tissue infections		Residual codes; unclassified		Other diseases of veins and lymphatic											
T18	42		14		18		15		33											
	Secondary malignancies		Cancer of colon		Cancer of other GI organs; peritoneum		Cancer of rectum and anus		Cancer of kidney and renal pelvis											
T19	109		38		45		98													
	Acute cerebrovascular disease		Non-Hodgkin's lymphoma		Maintenance chemotherapy; radiotherapy		Essential hypertension													
T20	103		27		257															
	Pulmonary heart disease		Cancer of ovary		Other aftercare															

proportions. The former is based on the maximum probability of some topics for the given patients while the proportions are calculated with the summation of posterior probabilities for each topic. This difference shows that some diagnosis code groups have more counts than others. Namely, for some diseases, patients have to pay more visits than other diseases. Hence, the patient topic distribution analyses can reveal the subtle nature of diseases.

Discussion and Limitations

Although many techniques, such as principle component analysis (PCA) [30], factor analysis (FA) [31] or probabilistic latent semantic indexing (pLSI) [32] have been used in clustering medical data, topic modeling has been proved to be a

model with distinct advantages. One of them is to group semantically related documents as well as terms together. In this work, LDA groups related diagnosis code groups into clusters. This provides strong interpretive potential in making phenotyping analysis or designing clinical decision support systems. Secondly, in contrast to PCA, FA or pLSI, LDA assume that each document may involve multiple components or topics and the generative process is based on Bayesian nature. Therefore, it is suitable for hierarchical analysis. Thirdly, the Dirichlet prior enables LDA can smooth its topic distribution, thus overcoming the overfitting problem of other models.

Another advantage is the unsupervised nature of LDA and its flexibilities. LDA itself does not require any training data or a priori knowledge about diseases. However, it doesn't prevent LDA to incorporate supervised information or external knowledge as prior or even as supervised labels. In our on-going work, one of our goals is to use section headers, physicians comments or labels on clinical notes as observed side information to train supervised or semi-supervised topic models for prediction tasks. LDA is designed for document analysis mainly because it is good at doing heterogeneous data analysis. Hence, it is now broadly used in image processing, bioinformatics and information retrieval. That is the main reason that we applied LDA in diagnosis code analysis.

Undoubtedly, there are limitations for the unsupervised LDA. The first limitation is the inconsistent mapping between the topics and the actual common properties of disease group. This can be found from the 20 topics generated. Some topics cluster some diagnosis code group together without much similarity. For example, *cancer of prostate* and *other liver diseases* in topic 16 seem not so related but they are the two highest code groups in it. Yet, we cannot say there is no reason for them to cluster together. They may be related due to some uncovered comorbidity. Finding out the exact cause requires addition information and domain knowledge. If we can add some supervised information, we may have a better control on the model generation and prediction. This may also imply that a topic is not necessarily associated with only one concept, and it could be related to several commonalities shared by diagnosis code group. The third limitation is that in this work, we didn't do much on the evaluations though we review and measure whether topics generated fit classification standard in AHRQ. It is still necessary to evaluate topics from other standards, such as similarity measurements, human judgments and so on.

In addition, there may be inconsistency for the results of each sampling. A common problem existing among sampling methods is its stability. LDA, starting with Dirichlet distributions, generates topic distributions. Next, it generates topics and diagnosis code groups in turn via a series of multinomial distributions. Although the conjugate nature between Dirichlet and multinomial distributions guarantee the theoretically soundness and the simplicity of the model, the results, after a few hundreds of iterations via Gibbs sampling, yielded are usually slightly different each time. Although we cannot fully control the stability of Gibbs sampling, Sato et al. [33] and Asuncion et al. [34] have proved that the collapsed variational Bayes inference with a zero-order Taylor expansion approximation, called CVB0 inference can get better performance than Gibbs sampling methods. Replacing the current inference methods with CVB0 can be one solution to explore in the future.

In this work, we identified 20 topics that could almost be connected with some group of diseases. However, we also observe that the same diagnosis code group might fall into different topics. For example, *Residual codes; unclassified* has been seen to share above 5% among 7 different topics. Based on the AHRQ definition, such codes cannot exactly be classified. This may be partially the reason that such codes are assigned to different topics. Such phenomenon is very popular in human languages considering the polysemy natures of words. But in diagnostic code grouping analysis, this may lead to confusions on the topic grouped together if we cannot find strong reasons for them. The phenomenon may need domain experts to interpret. Further distance assessment, like KL divergence or mutual information, may help find clearer demarks between each group.

Conclusions and Future Work

This study ivestigates the efficacy of topic modeling for the discovery of hidden patterns from a large epidemiology cohort. The results demonstrate that disease groups based on topic modeling do have statistically significance and also can reveal semantic commonalities among diseases. In our future work, we would add other patient information, such as drug, lab, procedure events and temporality to the analysis. In addition, temporal trends plays important roles in any epidemiological study. In addition, we would focus on an "interesting subpopulation" (e.g., a very complex or poorly understood disorder) to explore whether topic modeling help to unravel a complex disorder. The construction of temporal topic modeling on an epidemiology cohort may also lead to interesting discovery.

Acknowledgements

This work was made possible by joint funding from National Institute of Health R01LM009959A1 and National Science Foundation ABI:0845

References

1. Smith, C.A., *ElectronicHealth Records.* 2003.
2. Li, J.-s., H.-y. Yu, and X.-g. Zhang, *Data Mining in Hospital Information System.*
3. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet allocation.* Journal of Machine Learning Research, 2003. **3**: p. 993-1022.
4. Griffiths, T.L. and M. Steyvers, *Finding scientific topics.* Proceedings of the National academy of Sciences of the United States of America, 2004. **101**(Suppl 1): p. 5228-5235.
5. Bisgin, H., et al., *Mining FDA drug labels using an unsupervised learning technique-topic modeling.* BMC bioinformatics, 2011. **12**(Suppl 10): p. S11.
6. AlSumait, L., D. Barbará, and C. Domeniconi. *On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking*. in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. 2008. IEEE.
7. Xu, G., Y. Zhang, and X. Yi. *Modelling user behaviour for web recommendation using lda model*. in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*. 2008. IEEE.
8. Phan, X.-H., L.-M. Nguyen, and S. Horiguchi. *Learning to classify short and sparse text & web with hidden topics from large-scale data collections*. in *Proceedings of the 17th international conference on World Wide Web*. 2008. ACM.
9. Nguyen, C.-T., et al., *Web search clustering and labeling with hidden topics.* ACM Transactions on Asian Language Information Processing (TALIP), 2009. **8**(3): p. 12.
10. Bisgin, H., et al., *Investigating drug repositioning opportunities in FDA drug labels through topic modeling.* BMC bioinformatics, 2012. **13**(Suppl 15): p. S6.
11. Sauver, J.L.S., et al., *Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project.* American journal of epidemiology, 2011. **173**(9): p. 1059-1068.
12. Green, J., *The three C's of etiology.* Wide Smiles, 1996.
13. Loscalzo, J., I. Kohane, and A.-L. Barabasi, *Human disease classification in the postgenomic era: a complex systems approach to human pathobiology.* Molecular systems biology, 2007. **3**(1).
14. Bugrim, A., T. Nikolskaya, and Y. Nikolsky, *Early prediction of drug metabolism and toxicity: systems biology approach and modeling.* Drug discovery today, 2004. **9**(3): p. 127-135.
15. Cherkin, D.C., et al., *Use of the International Classification of Diseases (ICD-9-CM) to identify hospitalizations for mechanical low back problems in administrative databases.* Spine, 1992. **17**(7): p. 817-825.
16. Found, E.C., *ICD-9-CM Codes.*
17. Elixhauser, A., C. Steiner, and L. Palmer, *Clinical classifications software (CCS).* Book Clinical Classifications Software (CCS)(Editor ed^ eds), 2008.
18. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic acids research, 2004. **32**(suppl 1): p. D493-D496.
19. Hersh, W.R., et al. *TREC 2006 Genomics Track Overview*. in *TREC*. 2006.
20. Wang, H., et al., *Finding complex biological relationships in recent PubMed articles using Bio-LDA.* PLoS One, 2011. **6**(3): p. e17243.
21. Bellaachia, A. and E. Guven, *Predicting breast cancer survivability using data mining techniques.* Age, 2006. **58**(13): p. 10-110.
22. Jackson, S., et al., *Bacillus cereus and Bacillus thuringiensis isolated in a gastroenteritis outbreak investigation.* Letters in Applied Microbiology, 1995. **21**(2): p. 103-105.
23. Ogilvie, M.M. and C.F. Tearne, *Spontaneous abortion after hand-foot-and-mouth disease caused by Coxsackie virus A16.* British medical journal, 1980. **281**(6254): p. 1527.
24. Newman, D., S. Karimi, and L. Cavedon, *Using topic models to interpret MEDLINE's medical subject headings*, in *AI 2009: Advances in Artificial Intelligence*. 2009, Springer. p. 270-279.

25. Bimboim, H. and J. Doly, *A rapid alkaline extraction procedure for screening recombinant plasmid DNA*. Nucleic acids research, 1979. **7**(6): p. 1513-1523.
26. Chen, Y., et al., *A LDA-based approach to promoting ranking diversity for genomics information retrieval*. BMC genomics, 2012. **13**(Suppl 3): p. S2.
27. Chen, X., et al. *Inferring functional groups from microbial gene catalogue with probabilistic topic models*. in *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*. 2011. IEEE.
28. St Sauver, J.L., et al., *Data resource profile: the Rochester Epidemiology Project (REP) medical records-linkage system*. International journal of epidemiology, 2012. **41**(6): p. 1614-1624.
29. St Sauver, J.L., et al. *Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project*. in *Mayo Clinic Proceedings*. 2012. Elsevier.
30. Jolliffe, I., *Principal component analysis*. 2005: Wiley Online Library.
31. Kline, P., *An easy guide to factor analysis*. 1993.
32. Hofmann, T. *Probabilistic latent semantic indexing*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999. ACM.
33. Sato, I. and H. Nakagawa, *Rethinking collapsed variational Bayes inference for LDA*. arXiv preprint arXiv:1206.6435, 2012.
34. Asuncion, A., et al. *On smoothing and inference for topic models*. in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 2009. AUAI Press.

Electronic health records and disease registries to support integrated care in a health neighbourhood: an ontology-based methodology

Siaw-Teng Liaw[a,b,c], Jane Taggart[a], Hairong Yu[a], Alireza Rahimi[a]

[a]*University of New South Wales, Australia;* [b]*SW Sydney Local Health District, Australia,* [c]*Ingham Institute of Applied Medical Research*

Abstract

Disease registries derived from Electronic Health Records (EHRs) are widely used for chronic disease management (CDM). However, unlike national registries which are specialised data collections, they are usually specific to an EHR or organization such as a medical home. We approached registries from the perspective of integrated care in a health neighbourhood, considering data quality issues such as semantic interoperability (consistency), accuracy, completeness and duplication. Our proposition is that a realist ontological approach is required to systematically and accurately identify patients in an EHR or data repository of EHRs, assess intrinsic data quality and fitness for use by members of the multidisciplinary integrated care team. We report on this approach as applied to routinely collected data in an electronic practice based research network in Australia.

Keywords:

EHR, patient registries, data quality, routinely collected data, data repository, health neighbourhood, integrated care.

Introduction

Disease registries derived from Electronic Health Records (EHR) are widely used for chronic disease management (CDM). However, not enough is known about the quality of EHR-based registers in the UK (1, 2) and Australia (3). There are publications about large administrative or population health databases, but little about disease registries created from multiple EHRs. Even less information is available about whether improved quality of EHR-based disease registries improve CDM, patient safety or quality outcomes. In addition to research, the increasing use of EHR-based registries, created through "blackbox" extraction tools, for clinical care can increase the likelihood and scope of data errors and adverse events (4).

The design and development of EHR-based disease registries does not appear systematic or comprehensive (5). Aspects of quality of disease registries have been examined in the UK (2, 5) and through our own work on the consistency and quality of diabetes registries within an electronic Practice Based Research Network (ePBRN) in Australia (6).

Our proposition (7) is that a realist (8) and ontological (9) approach is required to systematically and accurately identify patients in an EHR (10), or data repository of information from multiple EHRs, and assess intrinsic data quality and fitness for use by stakeholders such as members of the multidisciplinary integrated care team or researchers (6). The realist approach (8) adopted for this evolving yet complex domain includes:

- **Context:** CDM, integrated care, evidence based practice;
- **Mechanisms:** systematic methods to assess and manage the quality of data integration, knowledge integration, clinical integration and interdisciplinary integration;
- **Impacts/outcomes:** improved data quality and fitness for use of disease registries, and, over the longer term, safety and quality of integrated care.

The ontological approach to EHR-based registers includes the collection of formal, machine-processable and human-interpretable representations of the entities, and the relations among those entities, within a defined domain (11). Ontologies also provide regimentations of terminology that can support the reusability and integration of data, thereby supporting the development of automated systems for data annotation, information retrieval, and natural-language processing (11). By incorporating defined rules, ontologies can generate logical inferences and control the inclusion/exclusion of relevant objects (12), such as the patient with a diagnosis of diabetes mellitus (DM), abnormal pathology (Path) test, DM medication (Rx), or a DM cycle of care Medicare service payment item (10). In addition, a formal ontological model of the domain data and metadata can specify a unified context which allows intelligent software agents to act in spite of differences in concepts and terminology from different primary care EHRs. This will enable the systematic development of automated, valid and reliable methods to extract, link and manage data as well as assess the data quality and semantic interoperability issues.

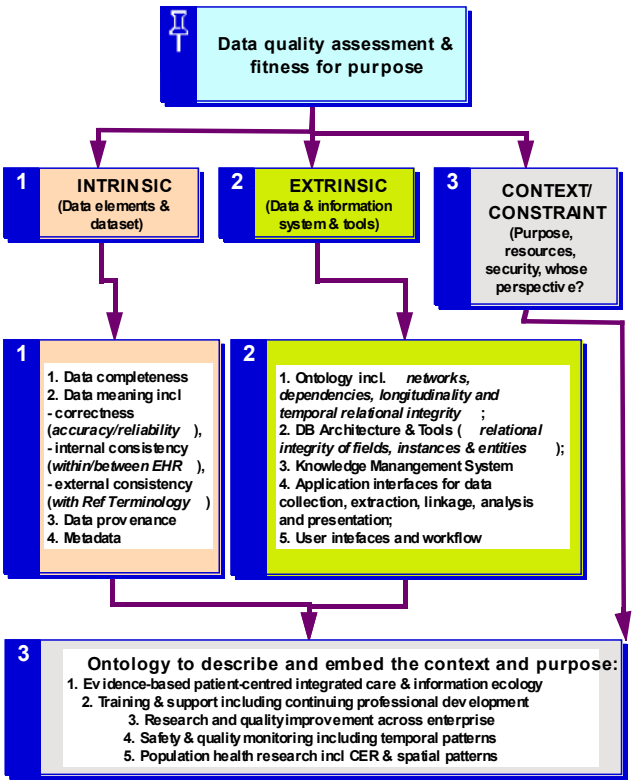
We have reported on our realist ontological approach ("Context-mechanisms-impact") to the quality of routinely collected data and integrated care, the relevant concepts and their relationships (13). The context is focused on the need for complete, correct, consistent and timely information about the cycle of care, risk factors, disease indicators, quality of life and patient satisfaction. The mechanism is the development and validation of ontologies to conceptualise and formalize the information and methods required to implement evidence-based integrated care in a range of contexts. This will allow the development of software agents to find cases to create disease registries, assess the intrinsic data quality and determine fitness for integrated care.

The quality of registries is influenced by the quality of EHR data, the case-finding system and associated quality processes, including currency and integrity, and the context such as clinical, insurance or other functions or objectives. Data quality

(DQ) is defined by the International Standards Organisation as: "the totality of features and characteristics of an entity that bears on its ability to satisfy stated and implied needs" (ISO 8402-1986, Quality Vocabulary). This "fitness for purpose/use"(14) definition is necessarily multidimensional requiring all intrinsic components and extrinsic associations of the entity to meet benchmarks and work together to achieve the purpose or meet the requirements.

An examination of the data quality literature (6, 15, 16) have led us to develop a more specific conceptual framework for data quality (DQ) and fitness for purpose (Figure 1).

Figure 1. Data quality & fitness for purpose framework



The framework comprises intrinsic, extrinsic and contextual dimensions, each with their concepts and relationships.

1. The intrinsic concepts cover the data elements and dataset, including the metadata, semantics (data meaning), provenance (who authored, where, when?) and constraints to the data meanings.
2. The extrinsic concepts cover the information system, including concept representation, ontology, temporal relationships system architecture and user interface.
3. The contextual determinants include the objectives of stakeholders such as the integrated care practitioner, resource constraints, security requirements, legislation, etc.

Data elements are assessed intrinsically in terms of consistency, correctness; data sets in terms of completeness and duplicate records (6). We are developing ontology-based tools to assess the information required to support integrated care in terms of timeliness and relational, historical and temporal integrity between concepts. Temporal and conceptual relationships may be dependent or independent factors. Relationships may be at a number of levels e.g. at the concept or table levels. The contextual determinants have been assessed qualitatively, aiming to guide clinical and organizational strategies to improve data quality to ensure fitness for purpose. The unified context will allow intelligent software agents to act in an environment of different concepts and terminology from different EHRs.

This paper will report and discuss this realist and ontological approach to developing automated, valid and reliable methods to define "cases" for a registry, manage data quality and determine fitness for purpose. We used the integrated care of diabetes mellitus in a health neighbourhood, as represented by the ePBRN, as a case study of the methodology of this work.

Materials and Methods

Setting: The ePBRN pilot group of 4 general practices has tested and validated the ePBRN data, processes and management in context, depending on the purpose. The internal validation of the ePBRN involved regular checking of the data and metadata using both automated and manual methods to examine the data repository. The data are also checked with probabilistic matching to assess the extent of duplicate patients and patients shared within the geographic region, the local health neighbourhood. The methodology was implemented with Microsoft SQL Server and an extension, Transact-SQL™ to link the server objects in the SQL Server with the heterogeneous datasets from multiple EHRs (17). The external validation of the ePBRN extraction tool involved a comparison against two other commercial data extraction tools (4).

Case-finding: The ePBRN ontological approach (10) used defined rules to generate logical inferences and control the inclusion/exclusion of the patient with a diagnosis of diabetes mellitus (DM), diabetes reason for visit (RFV), abnormal pathology (e.g. HbA1C, glucose tolerance test), diabetes medication (Rx) or glucose testing scripts, or a DM cycle of care item in the Medicare Benefit Schedule (MBS) (10). Following the query, the results were also analysed to exclude duplicate records/patients from the final result. This ontological approach was implemented and tested using SPSS and SQL, Each method acted as a control/validator for the other's accuracy. The benchmark was established with a manual examination of the results of SPSS and SQL queries on the smallest participating practice (Practice 1) contributing to the ePBRN data repository.

Data quality management: The conceptualization of the DQ ontology (Figure 1) included operationalising the reported core dimensions such as accuracy, currency and completeness (15) or completeness, correctness, consistency and timeliness (6, 16) and including duplicates (to account for aggregating multiple EHRs), temporal pattern (to account for the constantly changing clinical "big data") and timeliness which is important in integrated care. Validation of the conceptualization included discussions with practitioners and consumers of health care. The specification of the data quality ontology started with the definitions of completeness, consistency and correctness of data that we have reported previously (6).

Formalisation: To formalize the disease registry and DQ ontologies, we drew on the prevalent technical mechanisms and methodologies for ontology development, including knowledge acquisition, conceptualisation, semantic modelling, knowledge representation and validation (18, 19). Most used a layered approach (20) to incorporate clinical guidelines and rule-based approaches. The development tools used include: Protégé, a popular open source ontology editor and knowledgebase framework (http://protege.stanford.edu/); reference terminology (SNOMED-CT-Au); representation languages (Web Ontology Language (OWL), XML and RDF (Resource Description Framework)); query languages

(SPARQL Protocol and RDF Query Language); rules languages (Semantic Web Rule Language (SWRL)); logic ontology reasoners to provide automated support for reasoning tasks in ontology and instance checking through -ontopPro- (http://ontop.inf.unibz.it/), an ontology based data access (OBDA) application (21). The patient data, associated with instances of ontology classes or properties, is populated through -ontopPro-. The knowledge component of the infrastructure, related to conceptual terminologies defined by the specified ontology, was built using SNOMED CT-AU and Web Ontology Language (OWL: http://www.w3.org/TR/owl-features/) through Protégé. Details have been reported elsewhere (17) on how the RDF schema is mapped to logics to support formal semantics and reasoning. Formal semantics describes precisely the meaning of knowledge i.e. the semantics does not refer to subjective intuitions, nor is it open to different interpretations by different actors or machines (22). We used the layered ontology methodology to address semantic interoperability issues amongst different EHR in the ePBRN (23-27). This approach enables intelligent software agents to act in various semantic contexts in collaborative environments. We implemented and tested the DQ ontology, using SPSS and SQL tools, with the pilot ePBRN (N=95,056) data repository.

Results

Ontological approach to find cases for a diabetes registry

An overall prevalence rate of 2.8%, lower than expected for diabetes, was found for this pilot dataset. Table 1 shows data completeness of relevant indicators (RFV, Rx, Path) used for this paper and highlights that the ontological approach was more sensitive, finding more cases than a single database table query. The range of 0.2-4.8% for single factor and 1.1-5.7% for the ontological approach across practices, suggest that data quality is a significant factor. The pathology and medication tables contributed most. Case finding was improved, but the main limitation had been data quality dimensions like data completeness and consistency (5). The denominator was also important in assessing prevalence as some practices do not accurately represent active and inactive patients in the EHRs.

Table 1. Diabetes patients identified by diagnosis (RFV), HbA1C, medication, and ePBRN ontological approach

N = EHR flagged active patients	Practice 1 (N=3863)	Practice 2 (N=7028)	Practice 3 (N=23,162)	Practice 4 (N=30,717)	ePBRN (N=64,770)
Completeness of data:					
• All RFV (All DM RFV)	95% (4.3%)	87% (5.7%)	92% (4.9%)	99% (6.5%)	95% (5.8%)
• All Rx (All DM Rx)	80% (2.4%)	94% (8.4%)	96% (5.4%)	96% (6.6%)	95% (6.4%)
• All Path (HbA1C)	16% (0.8%)	61% (8.0%)	63% (1.3%)	66% (1.5%)	62% 2.4%)
• All 3 (RFV+Rx+Path)	82%	90%	90%	92%	90%
Diabetes indentified by:	N (%)	N (%)	N (%)	N (%)	N (%)
• Reason for visit (RFV)	37 (0.9)	231 (3.3)	387 (1.4)	787 (2.6)	1,442 (2.2)
• Diabetes medication	19 (0.5)	332 (4.7)	446 (1.9)	803 (2.6)	1,600 (2.5)
• HbA1c	8 (0.2)	334 (4.8)	468 (2.0)	809 (2.6)	1,619 (2.5)
• ePBRN ontological approach	43 (1.1)	403 (5.7)	602 (2.5)	1,042 (3.4)	2,090 (3.2)

Duplication and other dimensions of data quality

Table 2 shows up to 13% patient records matched across the participating EHR neighbourhood, suggesting that data quality assessment and management should include the extent of duplication of data with information sharing across the neighbourhood as well as within practices where there can be up to 3% duplication (Table 3). This has significance for clinical use of EHR data in integrated and shared care as well as secondary uses for research, population health and policy guidance.

Table 2. Record matching across general practices in a neighbourhood – shared patients

N=EHR active patients	**Pract 1 (N=3863)**	**Pract 2 (N=7028)**	**Pract 3 (N=23,162)**	**Pract 4 (N=30,717)**	**ePBRN (N=64,770)**
Practice (postcode)	**Records (%)**	**Records (%)**	**Records (%)**	**Records (%)**	**Records (%)**
Practice 1 (2176)		175 (2.5)	142 (0.6)	405 (13)	722 (1.1)
Practice 2 (2164)	173 (4.4)		327 (1.4)	691 (2.2)	1,191 (1.8)
Practice 3 (2171)	139 (3.4)	333 (4.7)		3,011 (9.8)	3,483 (5.4)
Practice 4 (2176)	400 (10)	692 (9.8)	3,005 (13)		4,097 (6.3)
Total	712 (18)	1200 (17)	3,474 (15)	4,107 (13)	9,493 (15)

Table 3. Record matching within general practices – duplicated records

Suburb (postcode)	EHR Active patients	Matched patients (%)	Matched records (%)
Practice 1 (2176)	3,863	10 (0.2%)	20 (0.5%)
Practice 2 (2164)	7,028	97 (1.3%)	198 (2.8%)
Practice 3 (2171)	23,162	220 (0.9%)	447 (1.9%)
Practice 4 (2176)	30,717	413 (1.3%)	830 (2.7%)
Total	64,770	740 (1.1%)	1,495 (2.3%)

Specifying and formalising the ontological approach

In addition to SQL tools, we have used the various ontology development tools mentioned to formalize the ontology work. The formal specification of the ontologies developed is available as Protégé files. Testing has being conducted with one of the participating practice (Practice 1) in the ePBRN, using – ontopPro- to map to the relational ePBRN data repository and implement the built-in reasoners. SPARQL and SWRL were used as the underlying query languages. However, this is the subject of another paper in preparation, which will also compare the utility and validity of SQL-based inductive versus ontology-based approaches and tools to create accurate patient/disease registries and assess/manage the quality of routinely collected data in the ePBRN data repository and its source EHRs.

Discussion

Research into the quality of routinely collected data in EHRs and EHR-based disease registries, especially in primary care, is an evolving field. While standards and benchmarks are being developed in this research domain, a realist and ontological approach is the most appropriate to understand what is being done in what context and with what impact, given that the processes and knowledge base are continually evolving, requiring ongoing monitoring, evaluation and reflection. The ePBRN research confirms this need to ground the research and development work in context and in the real world of health practice, where data is noisy and continually changing.

The ontological approach to case-finding identified a greater number of cases for inclusion in a disease/patient registry, highlighting the importance of this approach in the real world where data collection is suboptimal. Data quality management of aggregated information from multiple EHRs in a health neighbourhood to support integrated care must include the detection and management of duplicated records. Duplicates also lead to inaccurate public health and epidemiological research.

Ontologies deal with reality (**being**) and the transformation (**becoming**) of concepts as they interact with one another over time. An ontologically rich approach to the creation of patient registries from EHRs is essential to optimise accuracy (10). The effect of data quality is predictable as the disease registry is only as good as the EHR from which it is created – and there is much room for improvement in EHR data quality (6, 16). The improvement requires realist ecological approaches to the governance and provenance of data quality across the data cycle from collection to management to display and secondary use in other applications such as electronic decision support (16, 28). This approach recognises that the quality of electronic data collected as part of routine clinical practice is determined by more than just the GIGO – garbage in garbage out - principle. For instance, data models are influenced by the database management system, security and access management software, organisational processes for data collection and management, and the people in the organisation who enter and use data (4). The ePBRN foundational work reported here, along with others, has confirmed this to a significant extent.

As we validate the formal ontology tools developed in the ePBRN program and apply them to the development of fully automated methods to address the data quality of EHR and data repositories of ever increasing sizes, it is anticipated that this will build greater evidence for ontological approaches in the clinical and informational domains. The final tested ontologies and software tools can enable the systematic development of automated, valid and reliable methods to extract, link and manage data as well as assess/manage the data quality and semantic interoperability challenges.

Limitations:

This is a work in progress, evolving from a pilot phase to an established representative practice-based research network (and, given resources, a health information exchange to support evidence-based clinical practice). Having said that, the ePBRN foundational work has been systematic and robust in the methodology adopted:

1. to establish the ePBRN to reflect a local health neighbourhood with hospital, community health, general practice and other primary care services;
2. to refine and test the tools to extract, link and manage the data repository of routinely collected data in multiple EHRs; and
3. to make the transition from traditional management of "big data" from SQL and schematic relational databases to an ontological approach using semantic web principles and tools.

The data reported is neither representative nor timely; it is part of a pilot ePBRN to conduct our experiments to validate our methodologies with real world data from primary and secondary care settings. Our data across all projects shows that the quality of routinely collected data in EHRs is not only variable and suboptimal (6), but also continually evolving and changing with time. This emphasizes the need for cost-effective and validated automated methods to assess and manage data and information systems in a timely manner. The ePBRN program demonstrates that the challenge is great but surmountable.

Conclusion

The specification of a unified context to enable intelligent software agents to act, in spite of differences in concepts and terminology from different EHRs, will enable the systematic development of automated, relevant, valid and reliable me-

thods to extract, link and manage data as well as manage the data quality and semantic interoperability issues. This ontological approach to collecting, annotating, analysing and presenting clinical and scientific data is probably the only practical and sustainable solution to the information and data explosion. This is important to optimize the availability of good quality and relevant information to facilitate the safety and quality of integrated care as well as accurate and valid research.

Acknowledgments

The ePBRN research is supported in part by the University of New South Wales (UNSW) Major Research Equipment Infrastructure Initiative (2010, 2012), UNSW Medicine, Ingham Institute for Applied Medical Research and the Health Contribution Fund (HCF) Research Foundation (2013-14). We thank the participating general practices for their support and guidance, the conceptual input of Simon de Lusignan and Craig Kuziemsky, and contributions of co-researchers from the UNSW School of Public Health & Community Medicine and the following UNSW Research Centres: Primary Health Care & Equity, Health Informatics and Asia-Pacific Ubiquitous Health Care.

References

1. de Lusignan S, Sadek N, Mulnier H, Tahir A, Russell-Jones D, Khunti K. Miscoding, misclassification and misdiagnosis of diabetes in primary care. Diabet Med. 2012 2012 Feb;29(2):181-9.
2. Martin D, Wright J. Disease prevalence in the English population: a comparison of primary care registers and prevalence models. Soc Sci & Med 2009;68(2):266-74.
3. Ford D, Knight A. The Australian Primary Care Collaboratives: an Australian general practice success story. Med J Aust. 2010;193(2):90-1.
4. Liaw S, Taggart J, Yu H, de Lusignan S. Data extraction from electronic health records – existing tools may be unreliable and potentially unsafe. Aust Fam Physician. 2013;42(11):820-3.
5. Mehta A. The how (and why) of disease registers. Early Human Development. 2010;86(11):723-8.
6. Liaw S, Taggart J, Dennis S, Yeo A. Data quality and fitness for purpose of routinely collected data – a case study from an electronic Practice-Based Research Network (ePBRN). AMIA Annual Symposium 2011; Washington DC: Springer Verlag; 2011.
7. Liyanage H, Liaw S, Kuziemsky C, de Lusignan S. Ontologies to improve chronic disease management research and quality improvement studies – a conceptual framework. In: Aronsky D, Leong S, editors. Medinfo 2013; Copenhagen: Elsevier Press; 2013.
8. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review - a new method of systematic review designed for complex policy interventions. J Health Serv Res Policy. 2005;10(Suppl 1):21-34.
9. Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Human-Comput Stud. 1995;43(5-6).
10. de Lusignan S, Liaw S, Michalakidis G, Jones S. Defining data sets and creating data dictionaries for quality improvement and research in chronic disease using routinely collected data: an ontology driven approach BCS Informatics in Primary Care. 2011;19(3):127-34(8).
11. Rubin D, Lewis S, Mungall C, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. OMICS (Summer). 2006;10(2):185-98.
12. Perez-Rey D, Maojo V, Garcia-Remesal M, et al. ONTOFUSION: ontology-based integration of genomic and clinical databases. Comput Biol Med. 2006;36(7-8):712-30.
13. Liaw S, Rahimi A, Ray P, et al. Towards an ontology for data quality in integrated chronic disease: a realist review of the literature. . Int J Med Inform 2013;82(1):10–24.
14. Wang RY. A product perspective on total data quality management. Communications of the ACM. 1998;41(2 (Feb)):58-65.
15. Wang R, Strong D, Guarascio L. Beyond accuracy: what data quality means to data consumers. J Management Information Systems. 1996;12(4):5-33.
16. Liaw S, Chen H, Maneze D, et al. Health reform: is current electronic information fit for purpose? Emergency Medicine Australasia. 2011 Feb 2012;24(1):57-63.
17. Yu H, Liaw S, Taggart J, Rahimi A. Using Ontologies to Identify Patients with Diabetes in Electronic Health Records. Poster/demo. International Semantic Web Conference 2013; 2013; Sydney Australia: Springer-Verlag Berlin Heidelberg.
18. Kuziemsky C, Lau F. A four stage approach for ontology-based health information system design. Artificial Intelligence in Medicine 2010 2010;50:18.
19. Ying W, Wimalasiri J, Ray P, Chattopadhyay s, Wilson C. An Ontology Driven Multi-Agent Approach to Integrated e-Health Systems International Journal of E-Health and Medical Communications (IJEHMC). 2010 2010;1(1):12.
20. Colombo G, Merico D, Boncoraglio G, et al. An ontological modeling approach to cerebrovascular disease studies: The NEUROWEB case. Journal of Biomedical Informatics. 2010;43(4):469-84.
21. Rodríguez-Muro M, Calvanese D. Quest, a System for Ontology Based Data Access. KRDB Research Centre for Knowledge and Data, Free University of Bozen-Bolzano, 2012.
22. Dean M, Schreiber G, Bechhofer S, et al. OWL web ontology language reference. W3C Recommendation. 2004.
23. Chalortham N, Buranarach M, Supnithi T. Ontology Development for Type II Diabetes Mellitus Clinical Support System2009 3 March 2011. Available from: http://text.hlt.nectec.or.th/ontology/sites/default/files/CRdm2css_0.pdf.
24. Ganendran G, Tran Q, Ganguly P, Ray P, Low G. An Ontology-driven Multi-agent approach for Healthcare. HIC2002.
25. Ganguly P, Ray P, Parameswaran N. Semantic Interoperability in Telemedicine through Ontology-Driven Services. Telemedicine & e-Health. 2005;11(3):8.
26. Hadzic M, Chang E, editors. Ontology-based multi-agent systems support human disease study and control. International Conference on Self Organization and Adaptation of Multi-Agent and Grid Systems (SOAS); 2005 Dec 11; Glasgow, UK. Amsterdam, The Netherlands: IOS Press.
27. Hadzic M, Dillon DS, Dillon TS, editors. Use and Modeling of Multi-agent Systems in Medicine. Proceedings of the 20th International Workshop on Database and Expert Systems Application; 2009.
28. de Lusignan S, Liaw S, Krause P, et al. Key concepts to assess the readiness of data for International research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. IMIA Yearbook of Medical Informatics. 2011:112-21.

Address for correspondence

Professor Siaw-Teng Liaw

The General Practice Unit, Fairfield Hospital

PO Box 5, Fairfield, New South Wales 1860, Australia

Email: siaw@unsw.edu.au

Work phone: +61 2 96168520

Work fax: +61 2 96168400

Detailed Clinical Modelling Approach to Data Extraction from Heterogeneous Data Sources for Clinical Research

Sarah N. Lim Choi Keung, PhD[1], Lei Zhao, MSc[1], James Rossiter, PhD[1], Mark McGilchrist, PhD[2], Frank Culross, MSc, BSc[2], Jean-François Ethier, MD[3], Anita Burgun, MD, PhD[3], Robert A. Verheij, PhD[4], Nasra Khan, MSc[4], Adel Taweel, PhD[5], Vasa Curcin, PhD[6], Brendan C. Delaney, BM BCh, MD[5], Theodoros N. Arvanitis, DPhil[1]

[1]Institute of Digital Healthcare, WMG, University of Warwick, UK;
[2]Health Informatics Centre, University of Dundee, UK;
[3]INSERM UMR_S 872, France;
[4]NIVEL Netherlands Institute for Health Services Research, The Netherlands;
[5]Department of Primary Care and Health Sciences, King's College London, UK;
[6]Department of Computing, Imperial College London, UK

ABSTRACT

The reuse of routinely collected clinical data for clinical research is being explored as part of the drive to reduce duplicate data entry and to start making full use of the big data potential in the healthcare domain. Clinical researchers often need to extract data from patient registries and other patient record datasets for data analysis as part of clinical studies. In the TRANSFoRm project, researchers define their study requirements via a Query Formulation Workbench. We use a standardised approach to data extraction to retrieve relevant information from heterogeneous data sources, using semantic interoperability enabled via detailed clinical modelling. This approach is used for data extraction from data sources for analysis and for pre-population of electronic Case Report Forms from electronic health records in primary care clinical systems.

INTRODUCTION

One of the challenges in healthcare is the efficient reuse of routinely collected data for secondary purposes, such as clinical research. The main uses of electronic health records (eHRs) from patient registries or eHR systems in clinical research are for data analysis and for pre-population of electronic Case Report Forms (eCRFs). While existing patient records can sometimes fulfil all the requirements of a retrospective study analysis, the pre-population of eCRFs from eHRs can cover between 30% and 50% of the requirements[1], and integrated electronic data capture for eCRFs and eHRs can have an even higher overlap, depending on the study[2]. These highlight the potential of reusing clinical data while reducing the amount of redundant data entry (data recorded in clinical care that can be directly used for clinical research). Our research aims to support the interoperability between the clinical researcher tools and the clinical data within patient registries and eHR systems.

The TRANSFoRm project[3] aims to develop rigorous and generic methods for the integration of primary care clinical and research activities, to support patient safety and clinical research. The two clinical research support tools for researchers are the Query Formulation Workbench (QFW) and the eCRF Data Collection Tool. The QFW helps researchers to define studies with eligibility criteria sets for participants, build queries to identify eligible participants, flag patients, and extract data for analysis. The eCRF Data Collection Tool will support primary care practitioners to collect clinical study data and support the collection of patient reported outcome measures (PROMs) via web and mobile methods. In TRANSFoRm, the challenge is to bridge the gap between user requirements in terms of clinical study data items, and the execution of actual queries based on these requirements at the data sources. We adopt a two-level modelling approach[4-6] to separate out the more stable domain information from the various schema implemented by the heterogeneous data sources. The detailed clinical modelling (DCM) approach represents this accurately and will be described further in this paper.

The workflow and the involvement of the TRANSFoRm tools (specifically the QFW) and components are shown in Figure 1, from the definition of the study data extraction requirements to the actual queries at the data sources. In this paper, we focus on cohort identification. Taking the case of a researcher using the QFW to define a retrospective study of patients with Diabetes Mellitus, Step 1 involves defining the data to be extracted from the data sources, without needing to know the format or coding system used in individual data sources. In Steps 2 to 4, a number of TRANSFoRm components are involved to convert the data extract definition into semantically interoperable queries that can be executed at the respective data sources to return the requested data in the format defined by the user.

The remaining sections of this paper are structured as follows to describe DCM approach for semantic interoperability. The Methods section describes the DCM approach as a two-level modelling based on an information model and archetypes to constrain it. The Results section then demonstrates with examples how user requirements are mapped to a specific patient registry schema for data extraction. Finally, we discuss the use of the DCM approach in other TRANSFoRm tools, and finish with some conclusions and future work.

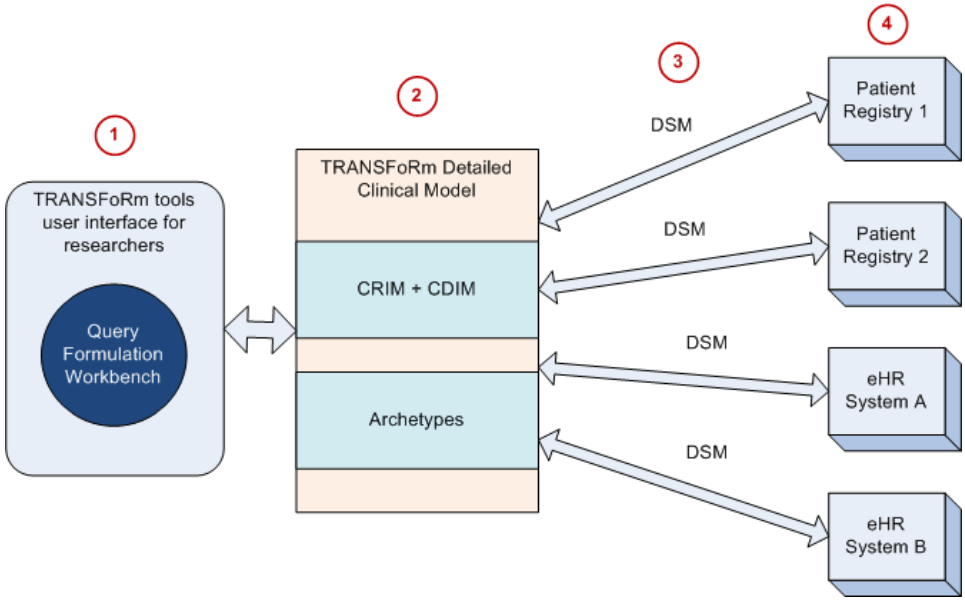


Figure 1. Conceptual workflow, from user definition of data extract requirements to actual queries at data source.

METHODS

Detailed Clinical Models (DCM) organise health information by combining knowledge, data element specification, relationships between elements, and terminology into information models that allow deployment in different technical formats[7,8]. DCM enables semantic interoperability by formalising or standardising clinical data elements which are modelled independently of their technical implementations. The data elements and models can then be applied in various technical contexts, such as eHR, messaging, data warehouses and clinical decision support systems. Work on DCM is still at an early stage with a number of groups involved on an ISO standard for DCM[9].

Within the TRANSFoRm project, the two-level modelling approach of DCM is depicted on the first level as an information model, the Clinical Research Information Model (CRIM), which defines the workflow and data requirements of the clinical research task, combined with the Clinical Data Integration Model (CDIM), an ontology of clinical primary care domain that captures the structural and semantic variability of data representations across data sources. This separation of the information model from the reference ontology has been previously described by Smith and Ceusters[10]. At the second level, archetypes are used to constrain the domain concepts and specify the implementation aspects of the data elements within eHR systems or patient registries. We use the Archetype Definition Language (ADL) to define the constraints and combine them with CDIM concepts in specifying the appropriate data types and range values. The two-level modelling approach, using the concept of archetype for detailed clinical content modelling, has been adopted by ISO/CEN 13606[11,12]. This approach makes it possible to separate specific clinical content from the software implementation. The technical design of the software is driven by the first level information model which specifies the generic information structure of the domain. The archetype defines the data elements that are required by specific application contexts e.g. different clinical studies.

The distributed query and data extraction infrastructure is a central component of the TRANSFoRm software platform. This infrastructure facilitates patient identification and reuse of routine healthcare data for research analysis. The TRANSForm platform interacts with disparate patient registries and eHR systems via the Data Node Connector, which translates the user queries, such as a data extraction definition as part of a retrospective study, in the form of archetypes to data source queries using the Semantic Mediator. The Semantic Mediator ensures the semantic translation queries from the Query Formulation Workbench to individual data source schema with the help

of data source models (DSM) and mappings to CDIM (CDIM-DSM)[13,14]. The transformed query can then be executed at the data source side and results are returned to the user. While specific DSM and CDIM-DSM mappings are required for each data source, these have to be built only once per data source. Additionally, the detailed clinical model is flexible enough to enable researchers to query heterogeneous datasets without any knowledge of the underlying structure, as they themselves do not use the DSM and CDIM-DSM mappings directly.

RESULTS

The data extraction for analysis was carried out for a Diabetes study, using a patient registry sample. In this section, we demonstrate how the data extract definition was processed, from the user at the Query Formulation Workbench, via the TRANSFoRm DCM to the data source. Following the steps in the conceptual workflow in Figure 1, we describe one specific data extract requirement – prescription dates for Metformin medication – for illustration. The clinical researcher defines what data to extract using the Query Formulation Workbench. In the case where the researcher wants to extract all the instances when patients have been prescribed Metformin (Figure 2), the data elements *Medication* and *Prescription date* are selected for extraction, and the constraint on the Medication concept is specified as part of the archetype specification. For example, the researcher can choose Metformin with the ATC code 'A10BA02' from the TRANSFoRm terminology service[15]. The resulting archetype definition in ADL is shown in Figure 3.

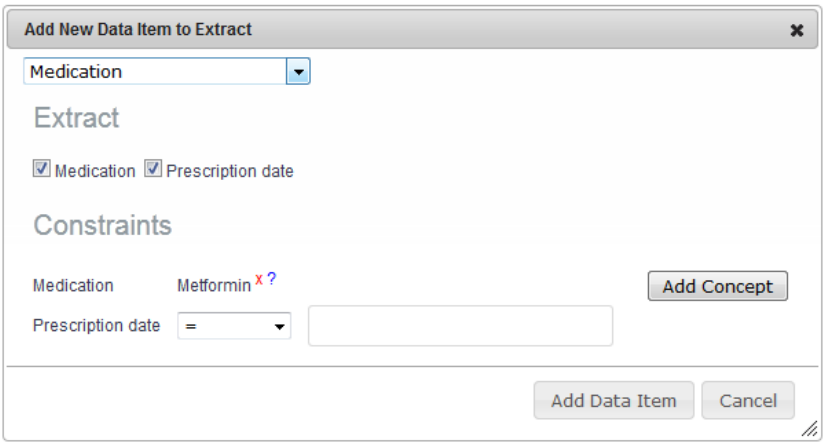


Figure 2: Data extract definition using the Query Formulation Workbench

```
TRANSFoRm-CRIM-ObservationResult.medication.v1.adl
    lifecycle_state = <"AuthorDraft">
    other_contributors = <>
    other_details = <>

definition
    ObservationResult[at0000] matches { -- Medication
        resultValue matches {   -- Medication code
            ELEMENT[at0001] {
                value matches { [ATC::A10BA02] }    -- Metformin
            }
        }
        effectiveTime matches { -- Prescription date
            ELEMENT[at0002] matches {
                value matches {*}
            }
        }
    }

ontology
    terminologies_available = <"CDIM", ...>
    term_bindings = <
        ["CDIM"] = <
            items = <
                ["at0001"] = <[CDIM::CDIM_000037]>    -- Formulated pharmaceutical data item
                ["at0002"] = <[CDIM::CDIM_000045]>    -- Written creation date
            >
        >
    >
```

Figure 3: Medication archetype definition in ADL.

The translation of archetypes into a computable form at the data source includes the use of a DSM (Figure 4a) and the CDIM-DSM mappings for the data source (Figure 4b). The DSM defines how the data source organises the medication prescription information, while the CDIM-DSM mappings express information in the form of triplets (CDIM concept; operator; terminology code). For instance, for Metformin with ATC code 'A10BA02', the information triplet is represented as (medication agent; =; 'A10BA02'). Following the transformations, an SQL query is generated to enable the specified data to be extracted from the data source (Figure 5).

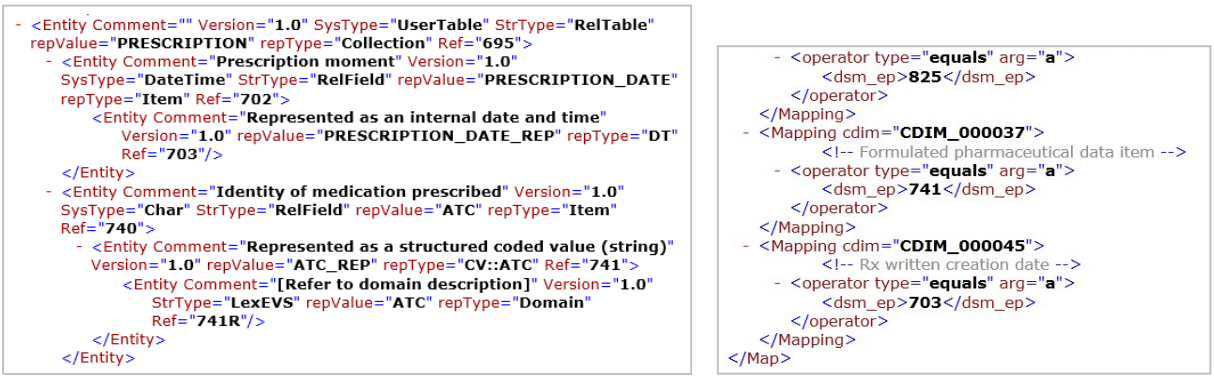


Figure 4: (a) Part of DSM definition **(b)** Part of CDIM-DSM for medication.

```
SELECT PATIENT.ID_PATIENT AS CDIM_000003, PRESCRIPTION.ATC AS CDIM_000037,
PRESCRIPTION.PRESCRIPTION_DATE AS CDIM_000045
FROM PRESCRIPTION
INNER JOIN PATIENT ON PRESCRIPTION.ID_PATIENT = PATIENT.ID_PATIENT
WHERE PRESCRIPTION.ATC IN ('A10BA02')
```

Figure 5: SQL query generated for data source schema.

DISCUSSION

Different solutions have been developed internationally to support a more rapid translation of scientific discoveries into clinical practice, notably i2b2[16]. i2b2 is a data warehousing system that extracts, transforms and loads data into a common schema. In comparison, the TRANSFoRm infrastructure adopts a model-based mediation approach, allowing the querying of heterogeneous data repositories without needing them to be in a single common schema. The TRANSFoRm project also aims to support clinical research with the reuse of eHR data within eCRFs, to avoid duplicate data collection. A minimisation of transcription errors and time-saving are added benefits for the reuse of routinely-collected clinical data. For instance, Köpcke et al.[17] report that the pre-population of case report forms decreased the time for data collection by nine-fold, from a median of 255 to 30 s. The DCM approach can be used in a similar way for the automatic pre-population of eCRFs from eHR systems as for the data extraction for retrospective studies from patient registries. The pre-populated data can be exported in the Operational Data Model (ODM) format[18], a standard for the interchange of data and metadata for clinical research, especially data collected from multiple sources. This will make the pre-populated data compatible with the remaining eCRF and PROM data that are collected as part of a study.

TRANSFoRm uses archetypes in the current implementation as ADL is a user-friendly language and can be easily understood by clinical researchers. HL7 templates, which constrain the HL7 clinical statement pattern, provide an alternative way to implement DCM in the context of HL7[8]. Future improvements to the TRANSFoRm GUI tools can include an authoring tool to assist users in defining new data elements. Referring to the medication archetype definition in Figure 2, currently, a user cannot directly update the archetype structure, for example to add the constraint of the dosage of the medication. Additionally, the tool can support various data element specification formats, such as HL7 templates and archetypes, for interoperation with systems that use these technologies.

CONCLUSION

The reuse of routinely collected data from clinical care in clinical research is an important goal of the TRANSFoRm project. The approach is to retrieve relevant data elements from the data sources (patient registries and eHR systems) without using a common structure to enable interoperability. Researchers can use the TRANSFoRm tools to define their studies without being aware of the underlying structure of the heterogeneous datasets. We have presented how a detailed clinical modelling approach is used to enable semantic interoperability between the researcher-defined queries and the individual data sources. The two-level modelling supports the flexibility of specifying new archetypes, as well as to add new data sources, while keeping the information model stable. Therefore, the DCM approach facilitates the bridging of the gap between clinical research and clinical care. The next steps include the validation of this approach and the related TRANSFoRm tools and components. Validation is being planned based on two use cases, a retrospective genotype-phenotype diabetes study and a prospective study for the gastro-oesophageal reflux disease randomised control trial.

Acknowledgements

The TRANSFoRm project is partially funded by the European Commission under the 7th Framework Programme (Grant Agreement 247787).

References

1. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P-Y, et al. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the EHR4CR platform. J Biomed Inform. 2011 Dec;44, Supplement 1:S94–S102.
2. Zahlmann G, Harzendorf N, Shwarz-Boeger U, Paepke S, Schmidt M, Harbeck N, et al. EHR and EDC Integration in Reality [Internet]. Appl. Clin. Trials. 2009 [cited 2013 Oct 1]. Available from: http://www.appliedclinicaltrialsonline.com/appliedclinicaltrials/article/articleDetail.jsp?id=641682
3. TRANSFoRm [Internet]. [cited 2013 Sep 30]. Available from: http://www.transformproject.eu/
4. Rector AL, Nowlan WA, Kay S, Goble CA, Howkins TJ. A framework for modelling the electronic medical record. Methods Inf Med. 1993 Apr;32(2):109–19.
5. Johnson SB. Generic data modeling for clinical repositories. J Am Med Inform Assoc. 1996;3(5):328–39.
6. Beale T. Archetypes: Constraint-based domain models for future-proof information systems. Seattle, Washington, USA, November 4, 2002; 2002. Available from: http://www.openehr.org/files/resources/publications/archetypes/archetypes_beale_oopsla_2002.pdf
7. Goossen W, Goossen-Baremans A, van der Zel M. Detailed Clinical Models: A Review. Health Informatics Res. 2010;16(4):201.
8. Goossen WTF, Goossen-Baremans A. Bridging the HL7 template - 13606 archetype gap with detailed clinical models. Stud Health Technol Inform. 2010;160(Pt 2):932–6.
9. European Committee for Standardization CEN. CEN/TC 251 - Standards under development [Internet]. [cited 2013 Oct 1]. Available from: http://www.cen.eu/CEN/Sectors/TechnicalCommitteesWorkshops/CENTechnicalCommittees/Pages/WP.aspx?param=6232&title=CEN%2FTC+251
10. Smith B, Ceusters W. HL7 RIM: an incoherent standard. Stud Health Technol Inform. 2006;124:133–8.
11. EN 13606 Association. The CEN/ISO EN13606 standard [Internet]. [cited 2013 Oct 2]. Available from: http://www.en13606.org/the-ceniso-en13606-standard
12. Muñoz P, Trigo J, Martínez I, Muñoz A, Escayola J, García J. The ISO/EN 13606 Standard for the Interoperable Exchange of Electronic Health Records. J Healthc Eng. 2011 Mar 1;2(1):1–24.
13. Ethier J-F, Dameron O, Curcin V, McGilchrist MM, Verheij RA, Arvanitis TN, et al. A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm. J Am Med Inform Assoc. 2013 Jan 9;20(5):986–94.
14. Ethier J, McGilchrist M, Burgun A, Sullivan F. D6.3 Data Integration Models [Internet]. 2013. Available from: http://transformproject.eu/TRANSFoRmproject.eu/Deliverables_files/TRANSFoRm%20D6%203%20Data%20Integration%20Models.pdf
15. Lim Choi Keung SN, Zhao L, Tyler E, Arvanitis TN. Integrated Vocabulary Service for Health Data Interoperability. Fourth International Conference on eHealth, Telemedicine and Social Medicine (eTELEMED 2012). Valencia, Spain: IARIA; 2012. p. 124–7.
16. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)[J]. J Am Med Inform Assoc. 2010, 17(2): 124-130.
17. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch H-U, et al. Secondary use of routinely collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. Int J Med Inf. 2013;82(3):185–92.
18. CDISC. ODM: Operational data Model [Internet]. [cited 2013 Oct 2]. Available from: http://www.cdisc.org/odm

Visualizing and Evaluating the Growth of Multi-Institutional Collaboration Based on Research Network Analysis

Jake Luo, PhD[1], Clara Pelfrey, PhD[2], Guo-Qiang Zhang, PhD[2]
[1]College of Health Science, University of Wisconsin Milwaukee;
[2]School of Medicine, Case Western Reserve University

Abstract

Research collaboration plays an important role in scientific productivity and academic innovation. Multi-institutional collaboration has become a vital approach for integrating multidisciplinary resources and expertise to enhance biomedical research. There is an increasing need for analyzing the effect of multi-institutional research collaboration. In this paper, we present a collaboration analysis pipeline based on research networks constructed from publication co-authorship relationship. Such research networks can be effectively used to render and analyze large-scale institutional collaboration. The co-authorship networks of the Cleveland Clinical and Translational Science Collaborative (CTSC) were visualized and analyzed. SciVal Expert™ was used to extract publication data of the CTSC members. The network was presented in informative and aesthetically appealing diagrams using the open source visualization package Gephi. The analytic result demonstrates the effectiveness of our approach, and it also indicates the substantial growth of research collaboration among the CTSC members crossing its partner institutions.

1. Introduction

Multi-institutional collaboration enhances the productivity and innovation of scientific research. Collaboration has been quickly changing the organization structure and research strategy of the biomedical research community[1-3]. Research collaboration network is a special type of social network within scientific communities. There has been a growing interest in analyzing the characteristics of collaboration network among research institutions. This creates an increasing need to evaluate the collaboration quality using network analysis methods[4]. Understanding the collaborative relationships among researchers and their affiliated institutions can help identify important network-based resources, such as leading members, rising personal, and strategic research clusters. Furthermore, collaboration network analysis can support the assessment and evaluation of research activity and productivity.

In biomedical science, organizations and leaders are also increasingly aware of the import roles of collaboration. Hence, developing efficient methods to objectively evaluate research collaboration becomes an important topic[5]. There have been many initiatives to develop new methods and theories for social network analysis (SNA)[6-9]. However, little work has been done to implement an efficient method for analyzing multi-institutional research collaboration network. In this paper we share our experience in developing a pipeline for research collaboration analysis[10], which not only provides quantitative measurement for decision-making, but also enables intuitive visualization of the key collaboration characteristics. The proposed framework uses co-authorship on scientific publications to generate a research network for collaboration analysis. The method is applied to analyzing the research collaboration of the Cleveland Clinical and Translational Science Collaborative (CTSC). The CTSC is among the early consortiums receiving NIH funding for the CTSA award[2]. The CTSC has been actively building collaborative infrastructure to support clinical and translational research for the five affiliated institutions, including Case Western School of Medicine (Case), Cleveland Clinic Foundation (CCF), University Hospital (UH), Metro Health, and Louis Stroke VA Medical Center.

In the next section, we describe the proposed method for transforming research publications to structured data sets for network analysis, followed by Results, Analysis and Discussions.

2. Method

The overall components and steps of the framework are illustrated in Figure 1. Our pipeline consists of four stages of information processing. The first stage "Information Extraction" (Figure 1) focuses on identifying relevant research documents and extracting author activities. A variety of documents can be used for research network analysis. Each type represents a specific aspect of collaborative activity. For example, multi-PI grant proposals indicate the sharing of complementary expertise and skills; clinical trial protocols show the collaboration on research project management; and publications reveal the co-authorship and imply the share of research responsibility and outcome. In this paper, we demonstrate the construction of a social network from the co-authorship data based on

scientific publications. We extract co-authorship data from affiliated research publications. Publication datasets are typically inexpensive and widely available to almost all research institutions, hence they are selected for this study.

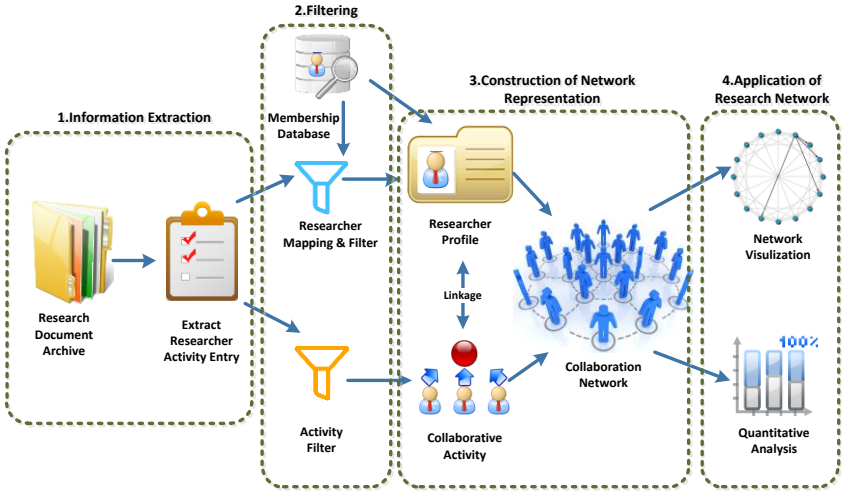


Figure 1: Systematic Research Network Generation

The second stage is "Mapping and Filtering," which focuses on preparing the extracted data for analysis. The documents retrieved from the first step normally contain information that is not relevant to network analysis. For example, non-affiliated researchers need to be filtered out. The best practice is to align the extracted researcher names with a membership database. In the alignment process, the names of the researchers will be disambiguated and mapped to their corresponding profile in the membership database, such as department, specialty. If a formal membership database does not exist, the process of disambiguation and profile alignment could be more challenging. Several prior studies proposed alternative methods for research profile alignment[11,12]. Another common filter is to limit the range of activities by specifying the year of publications or selecting a specific type of journals.

In the third stage, the social network is constructed and stored in a computable format. The previous filtering process results in two distinct types of data: the research profiles represent the entities of the research network, and the activity records (publications in our case) represent the relationships among the entities. Hence, it is essential to maintain the reference linkage of these two data sets during network construction. A researcher profile is represented as a "node" entity in the research network, while an activity record is transformed as one or more "edges" connecting the nodes. Two dataset tables are constructed and maintained for the nodes and edges respectively. A research collaboration network is then constructed by connecting the nodes (researchers) with their corresponding edges (collaboration activities). The constructed collaboration network can be used for quantitative analysis or rendered through visualization packages in the last step.

CTSC Research Network Construction

The publication data were extracted from SciVal Expert[TM13]. An XML parser was developed to extract the author information from the publication list. Since we focused on research collaboration among CTSC members in this study, non-CTSC members were filtered out by matching the author names to the CTSC membership database. CTSC researchers were represented as network graph nodes with their profiles assigned. Using the co-authorship list of the publications, we generated a pairwise coauthor list, which were used as edges to connect the nodes. To illustrate the interactions among research institutions, nodes (researchers) were colored by the affiliated institutions (CCF, Case Medical School, UH, MetroHealth, and VA center). The rendering of such multi-dimensional information in a compact and intuitive way is a challenge. We address this challenge using the force-directed graph algorithm and the open-source visualization package called Gephi[14]. The nodes are clustered by the Fruchterman Reingold algorithm[15] to show the members' connectivity power and similarity.

3. Results and Analysis

Research Network Visualization

Figure 2 (right) shows the research collaboration network of the CTSC based on 63,533 publications drawn from the SciVal database accumulated from 2008 to 2012. Figure 2 (left) shows the collaboration of 2008, which was the first year the CTSC was funded. Each node in the diagram represents a CTSC member. The names of the researchers are

removed in this paper for privacy reason. The color of a node represents the institution to which the member belongs. The size of a node shows the logarithmical connection degree. The larger the size, the more connections a member has. Connections are shown by the colored lines between nodes, with the color being assigned as that of the first author's affiliation. On the right, a network based on cumulative publications from year 2008 to 2012, shows that Cleveland Clinic (Red, 39%) and Case Medical School (Blue, 35%) represent the majority of the collaborative activities. University hospital (Green, 18%) also has a fair amount of collaborative members. MetroHealth Medical Center (Yellow, 6.18%) and the Louis Stokes Cleveland VA Medical Center (Brown, 0.94%) represent about 7 percent of the members. Comparing to the diagram on the left, the density of the nodes and edges has increased significantly, indicating substantial growth of collaboration among CTSC members across the partner institutions. Two independent evaluators examined the networks and confirmed the precision and the representativeness of the visual network. Note that some members solely collaborated within their own institutions, while others served as hubs that reached out to other research programs. Leaders of the institutions can be identified in the diagram by observing their strategic position in the diagram. The network also reveals researchers who were collaborative due to the possession of widely used services and technologies, such as Biostatistics.

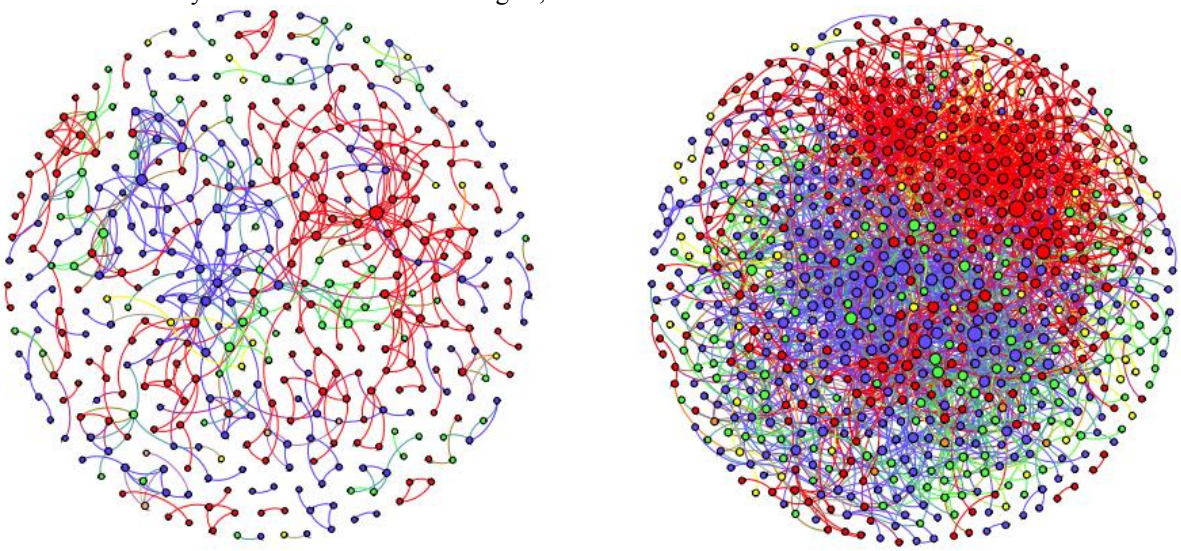


Figure 2: Left - collaboration network of the first year 2008; Right - collaboration network of 2008-2012

Figure 3 shows the cross-institutional collaboration during the years 2008, 2010 and 2012 respectively from left to right. The big circles delineate the five CTSC affiliated institutions. The color of the edges in Figure 3 is rendered with the combined colors of the two relevant institutions to help distinguish cross-institutional collaboration. For example, the edges between CWRU and CCF are in purple (a combination of blue and red), while the edges between CWRU and UH are in cyan (a combination of blue and green). The yearly network diagrams indicate that there has been a continuous growth of collaboration among the CTSC institutions.

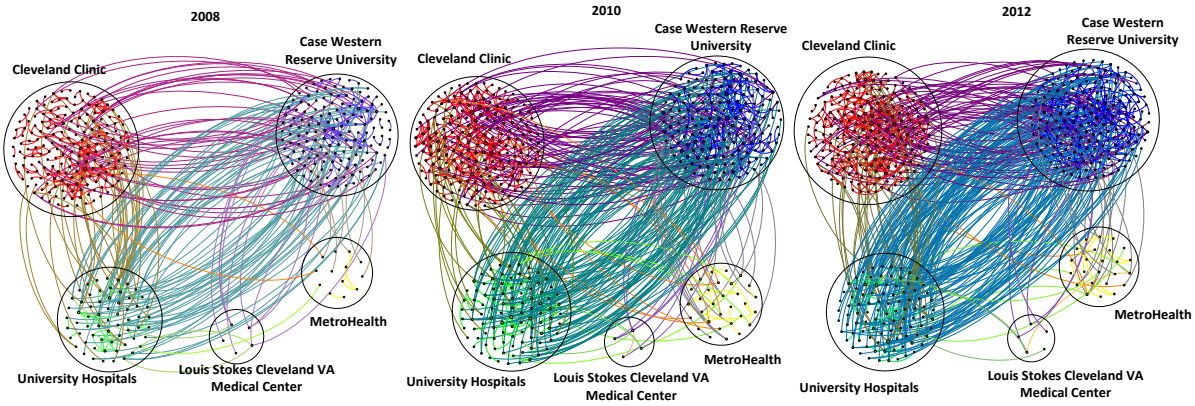


Figure 3: Growth of cross-institutional collaboration

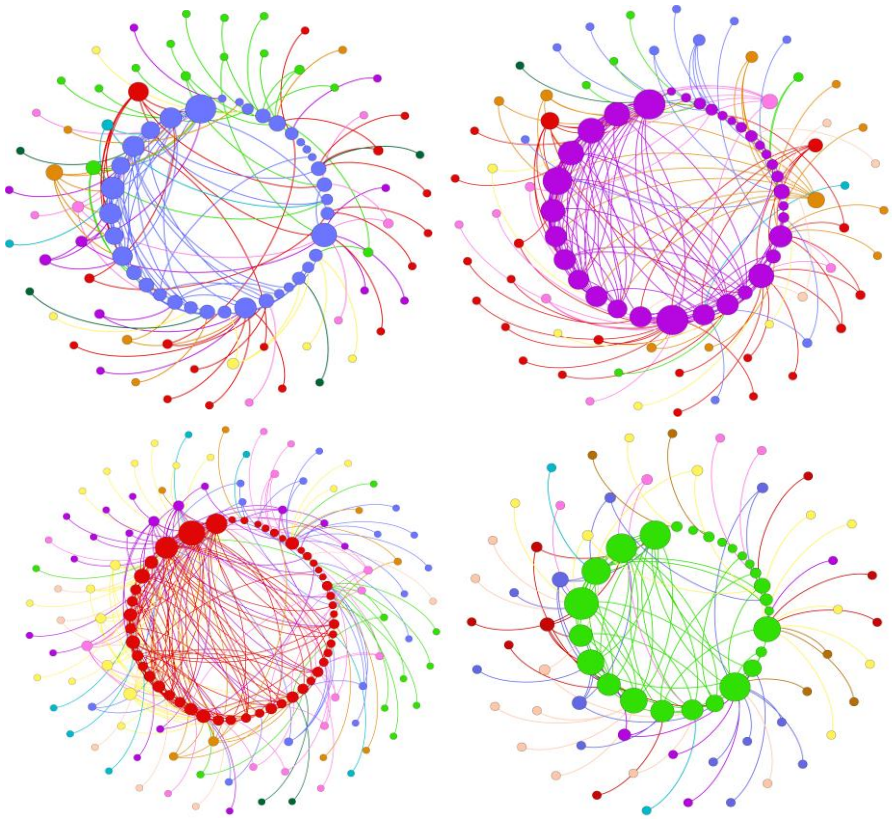


Figure 4: Network of individual scientific programs

Figure 4 shows the collaboration networks of individual scientific programs. The members of a program are shown in a circle. The color of the nodes in this figure represents the research program. Program members are sorted by their degree of intra-program co-authorships. The sorted sequence is arranged counter clockwise starting from 12 o'clock. Related inter-program connections are shown outside the main program circle.

Quantitative Analysis of Research Collaboration

To further quantify the growth of the CTSC research network across institutions, we analyzed the yearly percentage of cross-institutional collaborative publications and researchers. Table 1 shows the percentage of cross-institutional publications which were co-authored by researchers from two or more CTSC institutions. The publications are visualized as edges connecting the institutions in Figure 3. Cross-institution publications increased steadily at a 2%-3% rate each year from 2008 to 2012. In total, the collaborative publications increased 8.6%. Figure 5 (Left) shows the growth rate each year. Table 2 shows researchers who collaboratively published papers with other researchers from a different CTSC institution. The cross-institutional collaborative researchers are visualized as nodes in Figure 3. The result shows that the growth of collaborative researchers in CTSC was significant, from 24.9% to 61.1%. Figure 5 (Right) shows the total growth of researchers with collaborative publications. The results suggest that the CTSC is facilitating and promoting substantive research interactions among researchers from the affiliated institutions.

Table 1: Percentage of the cross-institution publications

Year:	2008	2009	2010	2011	2012
Cross-institution Publication	466	523	599	649	638
Total Publication	2909	2997	3019	3052	2589
Percent/Year	16.0%	18.0%	19.8%	21.3%	24.6%

Table 2: Researchers with cross-institutional publications

Year:	2008	2009	2010	2011	2012
Collaborative Researchers	177	306	399	461	515
Total Researcher	711	792	825	836	843
Percent/Year	24.9%	38.6%	48.4%	55.1%	61.1%

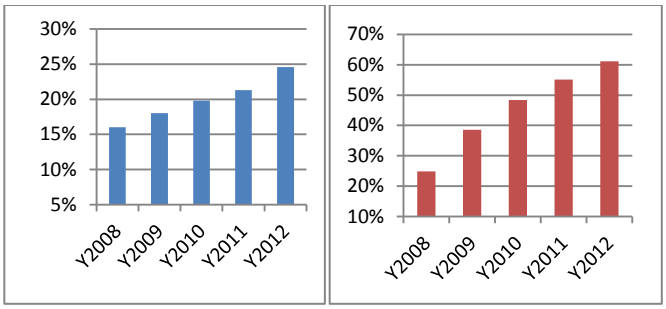


Figure 5: Left - Growth of cross-institutional publications; Right - Researchers collaborated to publish papers

4. Discussion

Many studies have discovered that a high level of research collaboration positively correlates with the quality and quantity of research outcomes[16-18]. An important strategic goal of the CTSA is to bridge the gap of biomedical research institutions, reduce barriers of communication[3], and increase the efficiency of collaboration between basic science researcher, clinical scientist and practicing physician. Hence, research collaboration is a key indicator for assessing the performance of a CTSA institution. In this CTSC case study, the network analysis results show a clear increasing trend of collaboration among the affiliated researchers. The overall quantity of the published paper also increased except for 2012. This may due to the lag of currency of information provided by SciVal Expert[TM].

Our network analysis pipeline provides an efficient method for evaluating cross-institutional collaborations. A bibliometric-based approached was used to extract co-authorship information for evaluating the collaboration. Although research publication co-authorship may not provide a comprehensive view of the collaboration process, it is still considered an effective and valuable information source for network analysis because of its advantages in availability and its faithful indication in research contribution[19-21]. In the biomedical research community, there are several ongoing efforts to build research networking tools and expert models to enable expertise discovery and research collaboration, such as Direct2Experts[4], CTSAconnect[22] and VIVO[23]. These platforms could provide additional data sources (e.g. facility usage record, clinical trials information) for network analysis. Our method complements these initiatives to provide an effective and self-contained pipeline to visualize and analyze the growth of multi-intuitional research collaboration.

Limitation

First, in this study we focused on analyzing the growth of research collaboration of the CTSC using the extracted publication data. Although many researchers may have external collaboration, the analysis was limited within the five CTSC affiliated institutions. To expand the analysis, we are expanding the data collection to other CTSA consortiums and planning to perform a large-scale network analysis for the CTSA collaboration. Second, social network analysis methods can be applied to measure other aspects of collaboration, such as individual researcher impact, connection diversity, and clustering degree. In the limited scope of this paper, we shared our results on developing an effective pipeline that transforms publication data into a suitable form for analyze the growth of research collaboration. The application scenario is highly desirable to many research institutions[24]. Hence, we believe our work provides an implementation blueprint and offers insights into the workflow of research collaboration analysis. In future work, we will expand the framework to provide more modules to assess the quality of research collaboration, such as analyzing the correlation between collaboration network and research output.

Conclusion

In this paper, we presented a streamlined pipeline for constructing research networks for collaboration analysis. Our pipeline is shown to be effective in supporting multi-institutional research network visualization and analysis. The approach enabled us to perform an objective evaluation to the research collaboration among the CTSC members using SciVal Expert[TM] data of 2008 to 2012. The results indicate that the collaboration has grown substantially since the inception of the CTSC. Not only the number of scientific publications shows substantial growth, the collaboration across the five partner institutions of the CTSC has increased.

Acknowledgements

We thank Mr. David Pilasky for providing the exported SciVal Expert™ data for this study.

This publication was made possible by the Clinical and Translational Science Collaborative of Cleveland, UL1TR000439 from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

1. Zerhouni EA. Clinical research at a crossroads: the NIH roadmap. *Journal of investigative medicine.* 2006;54(4):171-173.

2. Zerhouni EA. Translational and clinical science—time for a new vision. *New England Journal of Medicine.* 2005;353(15):1621-1623.

3. Woolf SH. The meaning of translational research and why it matters. *JAMA.* 2008;299(2):211-213.

4. Weber GM, Barnett W, Conlon M, et al. Direct2Experts: a pilot national network to demonstrate interoperability among research-networking platforms. *Journal of the American Medical Informatics Association.* October 28, 2011.

5. Greene SM, Hart G, Wagner EH. Measuring and Improving Performance in Multicenter Research Consortia. *JNCI Monographs.* November 1, 2005(35):26-32.

6. Scott J, Carrington PJ. *The SAGE handbook of social network analysis*: SAGE publications; 2011.

7. Merrill J, Hripcsak G. Using Social Network Analysis within a Department of Biomedical Informatics to Induce a Discussion of Academic Communities of Practice. *Journal of the American Medical Informatics Association.* November 1, 2008;15(6):780-782.

8. Park HW. Hyperlink network analysis: A new method for the study of social structure on the web. *Connections.* 2003;25(1):49-61.

9. Ellison NB, Steinfield C, Lampe C. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication.* 2007;12(4):1143-1168.

10. Luo Z, Sahoo SS, Zhang GQ. A Pipeline for Rendering and Analyzing Large Institutional Research Networks. *CTSA Informatics Key Function Committee meeting*. Chicago 2012; Page 44.

11. Han H, Giles L, Zha H, Li C, Tsioutsiouliklis K. Two supervised learning approaches for name disambiguation in author citations. Paper presented at: Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference.

12. Malin B. Unsupervised name disambiguation via social network similarity. Paper presented at: Workshop on link analysis, counterterrorism, and security 2005.

13. Vardell E, Feddern-Bekcan T, Moore M. SciVal Experts: A Collaborative Tool. *Medical reference services quarterly.* 2011;30(3):283-294.

14. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and manipulating networks. Paper presented at: ICWSM 2009.

15. Fruchterman TM, Reingold EM. Graph drawing by force-directed placement. *Software: Practice and experience.* 1991;21(11):1129-1164.

16. Katz JS, Martin BR. What is research collaboration? *Research Policy.* 1997;26(1):1-18.

17. Okubo Y, Sjöberg C. The changing pattern of industrial scientific research collaboration in Sweden. *Research Policy.* 2000;29(1):81-98.

18. Heinze T, Kuhlmann S. Across institutional boundaries?: Research collaboration in German public sector nanoscience. *Research Policy.* 2008;37(5):888-899.

19. Chin S-C, Madlock-Brown C, Street WN, Eichmann D. Firework visualization: A model for local citation analysis. Paper presented at: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium 2012.

20. Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. *Information processing & management.* 2005;41(6):1462-1480.

21. Barabâsi A-L, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications.* 2002;311(3):590-614.

22. Torniai C, Essaid S, Lowe B, Corson-Rikert J, Haendel M. Finding common ground: integrating the eagle-i and VIVO ontologies. *The Fourth International Conference on Biomedical Ontologies*. Montreal, Quebec 2013.

23. Krafft DB, Börner K, Corson-Rikert J, Holmes KL. The Future of VIVO: Growing the Community. *Synthesis Lectures on Semantic Web: Theory and Technology.* 2012:152.
24. Frechtling J, Raue K, Michie J, Miyaoka A, Spiegelman M. The CTSA National Evaluation Final Report. 2012.

Semi-Supervised Learning to Identify UMLS Semantic Relations

Yuan Luo[1], Ozlem Uzuner[1,2]
[1]Massachueets Institute of Technology [2]State University of New York at Albany

Abstract
The UMLS Semantic Network is constructed by experts and requires periodic expert review to update. We propose and implement a semi-supervised approach for automatically identifying UMLS semantic relations from narrative text in PubMed. Our method analyzes biomedical narrative text to collect semantic entity pairs, and extracts multiple semantic, syntactic and orthographic features for the collected pairs. We experiment with seeded k-means clustering with various distance metrics. We create and annotate a ground truth corpus according to the top two levels of the UMLS semantic relation hierarchy. We evaluate our system on this corpus and characterize the learning curves of different clustering configuration. Using KL divergence consistently performs the best on the held-out test data. With full seeding, we obtain macro-averaged F-measures above 70% for clustering the top level UMLS relations (2-way), and above 50% for clustering the second level relations (7-way).

Introduction
Biomedical documents are abundant in relations between concepts. For example, the sentence "The long-chain n-3 polyunsaturated fatty acids have cardioprotective effects, which may be partly due to their anti-inflammatory properties." states at least two relations: the "long-chain n-3 polyunsaturated fatty acids" *produces*[1] "cardioprotective effects"; the "cardioprotective effects" are *result_of* "their anti-inflammatory properties". However, such information, when locked in the narrative text, cannot be understood by computers due to lack of structure.

Mining relations from narrative text and making them accessible through a structured representation can benefit many studies, e.g., drug-drug, and drug-disease interaction studies. To this end, the Unified Medical Language System (UMLS) Semantic Network [1] can be used as a guide to align extracted structures. The network consists of the following:

- A set of semantic types, which provides a categorization of all concepts represented in the UMLS Metathesaurus®.
- A hierarchy of semantic relations, between semantic types.

This hierarchy has been developed and populated by hand in a top-down manner and undergoes periodic manual revisions [2]. As a result, collecting annotated relation instances and discovering those instances potentially characterizing new relations needs human annotation, which is labor intensive and time consuming. Our goal is to build a system that can expedite this process by mining relation instances from biomedical narrative text with little human intervention.

For this purpose, we need to be able to automatically identify the semantic relations between biomedical named entities. This is a nontrivial task as a semantic relation can be expressed in different ways using verbs (e.g., causes), prepositions (e.g., due to), or nouns (e.g., result of). For example, to say a symptom is the *result_of* a disease, one can use "due to", "caused by", or "result of". In addition, solely relying on keywords themselves can be problematic because a given word can be polysemantic. For example, the word "undergo" in "patient undergoes homeopathy procedure" indicates that the patient is *treated_by* a treatment. The same word in "patient undergoes a severe seizure" means that the patient *has_occurrence* a disease/symptom. Context of the word "undergo" is necessary to characterize the relations between the named entities.

The UMLS semantic relations are organized in a hierarchy. For example, the relations *disrupts*, *prevents*, and *complicates* are categorized under a more generic relation *affects*. We focus on the top two levels of this hierarchy and test our method for identifying relations at both levels. In the rest of this paper, we first review related work, then describe the construction of our corpus. After that, we explain our automatic relation extraction method in detail, and present experimental studies.

Related Work
Sematic relation extraction from biomedical narrative text is an active area of research. Rosario et al. [3] compared five graphical models and neural networks in classifying seven relations between diseases and treatments, where the neural network outperformed all graphical models. Plake et al. [4] used finite automata to learn from training samples. Khoo et al. [5] identified causal relations with manually created syntactic patterns from MEDLINE [6]. Sibanda et al. [7] used support vector machines (SVM) [8] to recognize disease-treatment relations in discharge summaries.

[1] Italic font in the main paper denotes the matching relations in the UMLS Semantic Network [1].

Clinical NLP systems such as MedLEE [9] and SemRep [10] apply hand-crafted syntactic and semantic rules to extract UMLS semantic relations. Co-occurrence patterns in MEDLINE [6] have also been explored to identify gene and protein synonyms [11], protein-protein interactions [12] etc. (see [13] for a review). Recently, semi-supervised or unsupervised acquisition of semantic relations has gained traction in the general NLP domain, where the methods typically include clustering and co-clustering algorithms that are often augmented with seeding or subsequent supervised classification [14][15][16]. We believe that these new developments towards demanding less annotated gold standard can shed light on the biomedical domain, where extracting the UMLS semantic relations largely depends on supervised learning.

Data Preparation

Our data set consists of the biomedical abstracts from the PubMed database [17]. We obtained the data set by crawling medical abstracts from the PubMed database that were returned in response to the query term "clinic". This query term was used to include a broad range of topics across the abstracts. We collected semantic entities mentioned in an abstract by applying the UMLS TFA parser [18] and extracting noun phrases from its phrase chunking output. We treated each noun phrase as a semantic entity and paired all phrases in one sentence. We filtered candidate semantic entity pairs based on whether a relation can exist between the semantic types of involved entities according to the UMLS Semantic Network. We focused on only the relations that are explicitly stated in the text. Two annotators, who have information science background and have completed college-level biology courses, annotated relations for each record. The annotators were presented with candidate semantic entity pairs and selected the best matching relations by following the UMLS semantic relation definitions. We found that some semantic relations in the third and fourth levels of the UMLS Semantic Network were either absent or were poorly represented in our corpus. Therefore, we limited ourselves to the seven relations in the top two levels of the UMLS Semantic Network. We performed double annotation for each semantic entity pair. The annotation lasted three months, covered 207 medical abstracts (3002 sentences) and produced 10082 semantic entity pairs. The initial Kappa statistic for inter-annotator agreement is 0.81 reflecting high agreement [19]. The annotators then discussed on disagreements and were able to resolve most of them. We discarded 124 pairs with irresolvable disagreements. The number of instances of each semantic relation in the gold standard data set is listed in Table 1.

Methods

We build a system that can automatically group UMLS concept pairs from biomedical narrative text into clusters where the grouping largely corresponds to the current semantic relation classes. We experiment with the k-means clustering framework under different configurations of distance metrics and seeding.

Relation	Count
Associated_with (AW)	9561
Spatially_related_to (SRT)	488
Functionally_related_to (FRT)	4719
Conceptually_related_to (CRT)	3177
Physically_related_to (PRT)	506
Temporally_related_to (TRT)	497
Isa (ISA)	397

Table 1 Semantic Relation distribution. AW and ISA are top level UMLS relations, the rest are in the second level. Note that there are 174 AW instances that do not fall in the second level *RT relations.

Figure 1 shows the workflow of our system. We use the UMLS TFA parser [18] to perform tokenization, part-of-speech tagging, and phrase chunking on narrative sentences. The phrases identified by the UMLS TFA parser constitute "minimal syntactic units" and consist of lexical elements. A lexical element can be a single-word term, or a multi-word term if that term is determined to be an independent unit in general English or medical dictionaries/thesauri (e.g., MeSH [20] and the UMLS SPECIALIST lexicon [21]). The TFA parser then applies semantic-syntactic rules over lexical elements to chunk them into phrases. For the resultant noun phrases, we extract their semantic types using UMLS MetaMap Transfer (MMTx) [22].

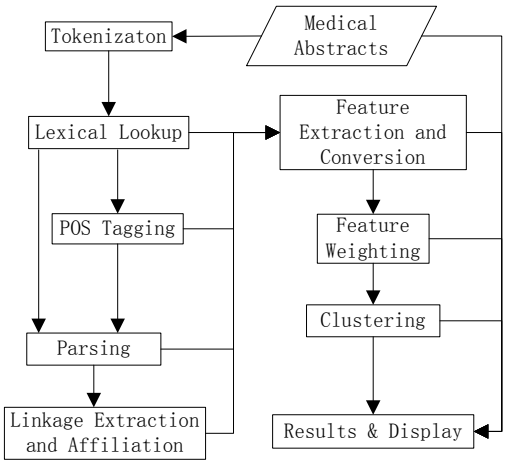


Figure 1 System workflow

In order to characterize the relation between the two semantic entities, our algorithm relies on the following features for that pair:

- Semantic features that include the UMLS semantic types of the phrases, e.g., "Quantitative Con-

cept" for "the high rate". This is motivated by the fact that certain semantic relation preferentially holds between specific sets of semantic types and vice versa.

- Lexical features that include all words in a sentence except for stop words. For example "high" and "rate" in the phrase "the high rate" in Figure 2. The intuition is that the words will help further distinguish semantic entities, in addition to semantic categories.
- Orthographic features that include punctuation, capitalization and the presence of digits. For example, capitalized phrase can be a proper name, often an instance of some disease, symptom etc. and likely to be in an *is_a* relation.
- Statistical features that include phrase length, sentence length and distance between phrases, all counted in terms of words. The intuition is that the relative distance between phrases can help differentiate semantic relations.
- Part-of-speech tags such as "verb" for "stain" and "noun" for "stain", which help distinguish that the word indicates a relation or is part of an entity.
- Syntactic features that include the syntactic links between semantic entities. Intuitively, similar link paths may indicate similar semantic relations, such as the two example link paths in Figure 2.

We next give more details on syntactic features. We use the Link Parser [23] to extract syntactic links between the words in a sentence. The Link Parser identifies 106 types of syntactic links and associates words with left and right connectors. A pair of compatible connectors forms a link. To adapt the Link Parser output for our task, we convert word links to phrase level links by performing the following steps: if the words at both ends of a link are in two different phrases, then that link is regarded as an inter-phrase link and is retained; if the words on both ends of a link are in one phrase, then the link is treated as an intra-phrase link and is discarded. For example, Figure 2 shows the result of applying the above procedures on the sentence "Further studies with more samples are needed in order to explain the high rate found among the pediatric patients in this research study". In Figure 2, we color multi-word phrases as blue and inter-phrase links as red. Other links are discarded.

We observe that the link labels often have semantic implications that are useful in characterizing the relations between the connected semantic entities. To explore this observation, we generalize Sibanda's syntactic n-grams [7]. We divide phrase-level links into introductory links, intermediate links, and closing links, according to whether they are before, between, or after the entity pairs. For example, in Figure 2, the semantic entity pairs "the high rate"-"the pediatric patients" (P1) and "the high rate"-"this research study" (P2) share the relation *occurs_in*. These two pairs have common link types as well. Tracing the intermediate link path for the pair P2 as highlighted with green in Figure 2, the intermediate links include "Mv" (indicating participle modifiers), "MVp" (connecting verbs to modifying preposition phrases), and "Js" (connecting prepositions to their objects). A similar analysis shows that the pair P1 shares the same intermediate links (highlighted with yellow in Figure 2) once ignoring subscripts[2], suggesting that similar relation holds here as in P2.

We construct link bigrams for all three types of links. Motivated by the observation that longer link span may lead to weaker phrasal relation, we also add the link span (defined as 1 + #phrases between link endpoints) as a feature.

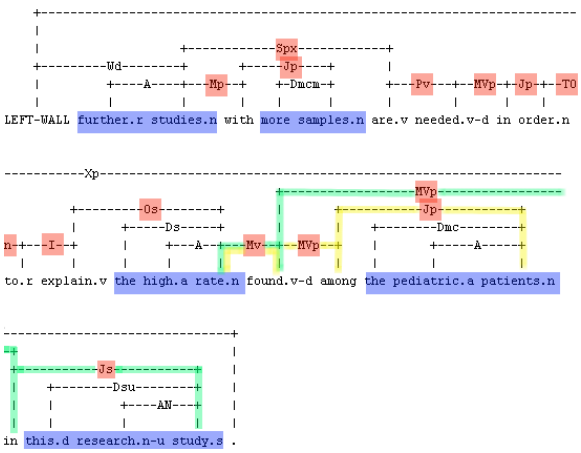


Figure 2 Link grammar output of an example sentence. In this example, multi-word phrases are highlighted with blue color and inter-phrase links are highlighted with red color. Two example link paths are also highlighted with green and yellow respectively.

Clustering Semantic Relations

We use the k-means clustering algorithm as we already know the number of clusters. Denote the data set as $Y = \{y_1, \ldots, y_n\}$, we want to form k disjoint clusters $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_k\}$, that is, $\cup_{i=1}^{k} \hat{y}_i = Y$, and $\forall i \neq j,\ \hat{y}_i \cap \hat{y}_j = \emptyset$. Let $X = \{x_1, \ldots, x_w\}$ be the features. These features are transformed into appropriate distance metrics between data points, which guide the formation of clusters by k-means. We use the Gmeans package [24], which minimizes an aggregated measure of intra-cluster distances called incoher-

[2] Subscripts are used to encode fine grammatical constraints. For example, the "p" is a subscript in "MVp", indicating prepositional modifying phrases to verbs.

ence. We experiment on the seeded k-means with several distance metrics including the Euclidean distance, the cosine similarity, and the Kullback-Leibler (KL) divergence. Let c_j be the center of the cluster j, which is computed by averaging across all data points in that cluster. The Euclidean distance incoherence is

$$\mathcal{E}(\{\hat{y}_j\}_{j=1}^k) = \sum_{j=1}^k \sum_{y \in \hat{y}_j} (y - c_j)^T (y - c_j).$$

The cosine similarity incoherence is calculated as $\mathcal{Q}(\{\hat{y}_j\}_{j=1}^k) = \sum_{j=1}^k \sum_{y \in \hat{y}_j} y^T c_j$.

The KL divergence incoherence indicates the information loss due to clustering and is formulated as

$$\mathcal{D}(\{\hat{y}_j\}_{j=1}^k) = \sum_{j=1}^k \sum_{y \in \hat{y}_j} p(y) KL\left(p(X|y), p(X|\hat{y}_j)\right)$$

where $X, y, \hat{y}_j$ are viewed as random variables.

Experimental Results and Discussions

We evaluate our system on the top two levels of the UMLS Semantic Network. To test the generalizability of our semi-supervised clustering, we split our corpus into a training set and a testing set at a 8:2 ratio, stratified by semantic relation types at the second level (top level then automatically stratified). For each distance metric and seeding configuration, we run k-means clustering 30 times to obtain statistically robust results. For each run, we randomly draw seeds at specified fraction. We evaluate performance by creating a confusion matrix; we assign cluster labels so that we can obtain the confusion matrix with the strongest diagonal [25]. We then compute per-class as well as micro- and macro- averaged precision, recall and F-measure, which are common clustering evaluation metrics [26]. Let TP denote the number of true positives, FP denote the number of false positives and FN denote the number of false negatives, the definition of precision is $P = TP/(TP + FP)$, recall is $R = TP/(TP + FN)$, F-measure is $F = 2 \times P \times R/(P + R)$. For each distance metric-seeding configuration, we report averaged results over the 30 runs. We find that the KL divergence consistently gives the best F-measures for varying seed fractions.

Figure 3 shows the learning curves of the seeded clustering with KL divergence as the distance metric. For both levels of the UMLS semantic relation hierarchy, the performance keeps improving beyond 50% seeding, but with decreasing speed. Due to space limitation, we leave the results of the other two distance metrics in the Appendix B, which were inferior to those of the KL divergence. This suggests that the seeded clustering is sensitive to the choice of the distance metric and that the KL divergence is a suitable distance metric on our dataset. See Appendix B for more comparisons.

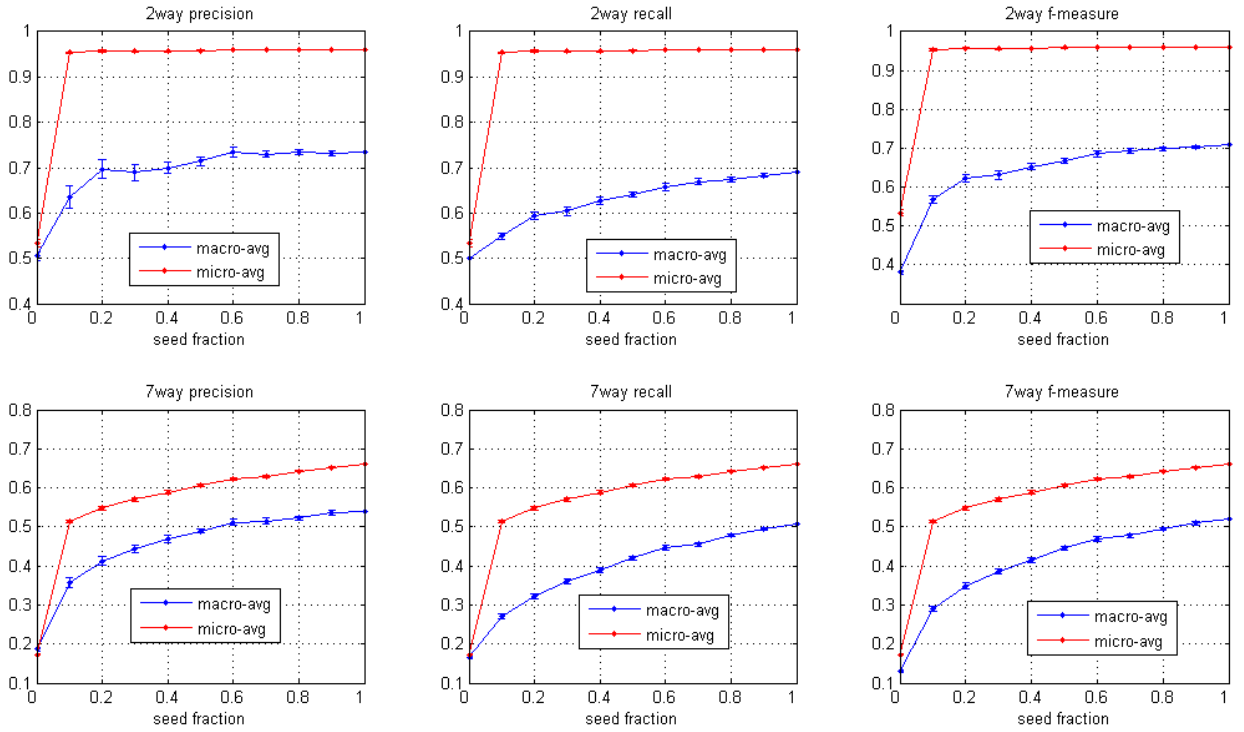


Figure 3 Macro-averaged and micro-averaged Precision, Recall and F-measure on 2-way and 7-way relation using KL divergence as the distance metric. Results are averaged over 30 runs, confidence intervals at $\alpha = 0.05$ are also shown, most of which are small, suggesting statistical stability.

With full seeding on the training data, our method generates a 2-way macro-averaged F-measure above 70% and a 7-way macro-averaged F-measure above 50%. With half seeding, our method still achieves a 2-way macro-averaged F-measure above 65% and a 7-way macro-averaged F-measure around 45%. This result suggests that the demand for seeding can be relaxed and our method can be used to automatically identify relation instances and provide a reasonable starting point with pre-labels to facilitate the human annotation process.

We present detailed per-class evaluation on k-means algorithm using KL divergence with varying seed fractions (up to 50%[3]) in Table 2 and Table 3. Not surprisingly, we see that the class imbalance took its toll on the clustering performance. For example, in the top level (Table 2), we have 9561 examples of the AW relation vs. 397 examples of the ISA relation. Precision and recall of ISA is much lower than AW. For the second level (Table 3), big performance drop is also seen in less populated classes such as AW[4], PRT and TRT.

Table 2 Performance per class of k-means clustering using KL divergence with random seeds on the relations from the top level of the UMLS Semantic Network.

Seed	Relation	Precision	Recall	F-measure
10%	AW	96.45%	98.83%	97.62%
	ISA	30.78%	10.94%	15.78%
20%	AW	96.79%	98.83%	97.80%
	ISA	42.62%	19.83%	26.64%
30%	AW	96.88%	98.61%	97.74%
	ISA	41.09%	22.31%	28.54%
40%	AW	97.05%	98.50%	97.77%
	ISA	42.84%	26.75%	32.64%
50%	AW	97.16%	98.55%	97.85%
	ISA	45.81%	29.53%	35.78%

Conclusion and Future Work

We presented a semi-supervised approach to automatically identify semantic relations according to the definitions in the UMLS Semantic Network. We created a corpus of semantic entity pairs whose relations were doubly annotated according to the top two levels of the UMLS semantic relation hierarchy. We demonstrated that our semi-supervised method has reasonable accuracy and coverage at both levels of resolution. By studying the learning curves of the seeded k-means with the KL divergence as the distance metric, we showed that the demand for seeding in the training data can be relaxed by half without greatly decreasing the performance. Therefore, our system can be used to assist with expert reviews on the semantic relation annotation task.

For future work, we note that our seeding is random and does not take into consideration how informative an example is. An active learning approach for picking the seeds could potentially further reduce the amount of required seeds and maintain similar levels of precision, recall, and F-measure.

Table 3 Performance per class of k-means clustering using KL divergence with random seeds (fractions 10% to 50%) on the relations from the second level of the UMLS Semantic Network. AW here includes AW instances that do not fall into *RT categories.

Seed	Relation	Precision	Recall	F-measure
10%	AW	27.50%	8.43%	12.20%
	CRT	49.35%	55.02%	51.93%
	FRT	57.92%	64.77%	61.12%
	ISA	36.39%	15.68%	21.47%
	PRT	21.04%	10.83%	14.09%
	SRT	35.90%	21.31%	26.39%
	TRT	22.38%	13.71%	16.73%
20%	AW	33.56%	16.96%	21.79%
	CRT	54.12%	58.30%	56.09%
	FRT	60.41%	67.25%	63.63%
	ISA	43.84%	18.76%	26.13%
	PRT	26.91%	15.17%	19.21%
	SRT	44.19%	30.79%	36.05%
	TRT	26.21%	18.61%	21.50%
30%	AW	34.11%	16.76%	21.54%
	CRT	56.80%	59.90%	58.28%
	FRT	62.63%	68.56%	65.44%
	ISA	47.86%	25.09%	32.65%
	PRT	30.22%	19.57%	23.57%
	SRT	50.84%	38.69%	43.78%
	TRT	28.47%	23.20%	25.35%
40%	AW	36.77%	20.39%	25.41%
	CRT	59.03%	63.00%	60.94%
	FRT	64.05%	68.99%	66.42%
	ISA	52.89%	29.44%	37.65%
	PRT	34.19%	22.57%	27.01%
	SRT	51.68%	43.30%	47.01%
	TRT	29.86%	23.98%	26.47%
50%	AW	36.15%	23.43%	28.01%
	CRT	60.55%	65.17%	62.77%
	FRT	66.11%	69.55%	67.78%
	ISA	55.49%	33.80%	41.86%
	PRT	37.13%	24.73%	29.58%
	SRT	55.04%	49.97%	52.26%
	TRT	32.47%	27.28%	29.54%

[3] Results for seed fraction above 50% are shown in Appendix A.

[4] In the second level, AW refers to those 174 instances that do not fall into *RT relations.

References

1. UMLS Semantic Network http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html
2. McCray AT. An upper level ontology for the biomedical domain. Comp Funct Genom 2003; 4:80-4.
3. Rosario B, Hearst M. Classifying Semantic Relationships in Bioscience Text. ACL 2004.
4. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. Ali Baba: PubMed as a graph. Bioinformatics, 22(19): 2444-2445, 2006
5. Khoo C, Chan S, Niu Y. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. ACL 2000. pp. 336-343.
6. MEDLINE http://www.nlm.nih.gov/pubs/factsheets/medline.html
7. Sibanda T. Was the Patient Cured? Understanding Semantic Categories and Their Relationships in Patient Records. Master Thesis, MIT, 2006.
8. Vapnik V. The Nature of Statistical Learning Theory. Berlin : Springer-Verlag, 1995.
9. Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. Exploiting semantic relations for literature-based discovery. AMIA 2006
10. Rindflesch, T and Marcelo, F. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics 36.6 (2003): 462-477.
11. Cohen, A. M., Hersh, W. R., Dubay, C., & Spackman, K. (2005). Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC bioinformatics, 6(1), 103.
12. Bunescu, R., Mooney, R., Ramani, A., & Marcotte, E. (2006, June). Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline. In Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis (pp. 49-56). Association for Computational Linguistics.
13. Cohen, T., & Widdows, D. (2009). Empirical distributional semantics: methods and biomedical applications. Journal of Biomedical Informatics, 42(2), 390-405.
14. Bollegala, T, Yutaka M, and Mitsuru I. "Relational duality: Unsupervised extraction of semantic relations between entities on the web." Proceedings of the 19th international conference on World wide web. ACM, 2010.
15. Sun, A, Ralph G, and Satoshi S. "Semi-supervised Relation Extraction with Large-scale Word Clustering." ACL. 2011.
16. Mohamed, T, Estevam H, and Tom M. "Discovering relations between noun categories." Proceedings of EMNLP, 2011.
17. PubMed. NCBI. http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed
18. Text Tools from Lexical System Group. http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/textTools/current/Usages/Parser.html.
19. Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions, Wiley.
20. MeSH: http://www.ncbi.nlm.nih.gov/mesh
21. Specialist Lexicon. http://lexsrv3.nlm.nih.gov/LexSysGroup/index.html.
22. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA 2001.
23. Sleator D, Temperley D. Parsing english with a link grammar. Technical Report, Carnegie Mellon University. 1991.
24. Dhillon I, Guan Y. Information-Theoretic Clustering of Sparse Co-Occurrence Data. UTCS Technical Report #TR-03-39.
25. Dhillon IS, Guan Y. Information Theoretic Clustering of Sparse Co-Occurrence Data. Proceedings of the Third IEEE International Conference on Data Mining, p.517, November 19-22, 2003.
26. Manning, C.D. et al. 2008. Introduction to information retrieval. Cambridge University Press Cambridge.

Appendix A

This section continues the per-class evaluation for k-means using KL divergence. Table 4 shows the continuation of Table 3 with seed fraction increasing to 100%. Table 5 shows the continuation of Table 2 with seed fraction increasing to 100%.

Table 4 Performance per class of k-means clustering using KL divergence with random seeds (fractions 60% to 100%) on the relations from the second level of the UMLS Semantic Network. See Table 1 for notation on abbreviations.

Seed	Relation	Precision	Recall	F-measure
60%	AW	39.19%	25.69%	30.38%
	CRT	62.59%	66.51%	64.48%
	FRT	67.42%	70.53%	68.93%
	ISA	55.60%	35.47%	43.18%
	PRT	40.10%	28.87%	33.46%
	SRT	58.65%	56.80%	57.57%
	TRT	34.28%	28.74%	31.16%
70%	AW	40.30%	27.06%	31.80%
	CRT	63.34%	68.07%	65.61%
	FRT	68.46%	70.49%	69.45%
	ISA	56.28%	36.71%	44.36%
	PRT	39.91%	29.17%	33.64%
	SRT	58.41%	58.45%	58.35%
	TRT	33.87%	29.46%	31.44%
80%	AW	35.76%	26.08%	29.75%
	CRT	64.35%	69.26%	66.71%
	FRT	70.13%	70.96%	70.54%
	ISA	56.96%	39.19%	46.36%
	PRT	42.92%	32.83%	37.16%
	SRT	61.37%	65.22%	63.17%
	TRT	34.92%	31.09%	32.86%
90%	AW	38.06%	28.33%	31.95%
	CRT	65.31%	70.87%	67.97%
	FRT	71.43%	71.20%	71.31%
	ISA	58.41%	41.15%	48.26%
	PRT	44.52%	35.70%	39.60%
	SRT	62.01%	67.66%	64.68%
	TRT	34.89%	31.33%	33.00%
100%	AW	34.48%	29.41%	31.75%
	CRT	66.09%	72.24%	69.03%
	FRT	72.71%	71.40%	72.05%
	ISA	58.93%	42.31%	49.25%
	PRT	45.00%	36.00%	40.00%
	SRT	63.89%	71.13%	67.32%
	TRT	35.87%	33.67%	34.74%

Table 5 Performance per class of k-means clustering using KL divergence with random seeds on the relations from the top level of the UMLS Semantic Network. See Table 1 for notation on abbreviations.

Seed	Relation	Precision	Recall	F-measure
60%	AW	97.29%	98.62%	97.95%
	ISA	49.54%	32.78%	39.28%
70%	AW	97.39%	98.46%	97.92%
	ISA	48.66%	35.47%	40.94%
80%	AW	97.43%	98.48%	97.95%
	ISA	49.59%	36.41%	41.95%
90%	AW	97.49%	98.36%	97.92%
	ISA	48.77%	38.16%	42.79%
100%	AW	97.55%	98.32%	97.94%
	ISA	49.21%	39.74%	43.97%

Appendix B

Figure 4 shows the learning curves of using the cosine similarity as the distance metric in clustering semantic entity pairs according to the top two levels of UMLS semantic relations. Figure 5 shows the corresponding learning curves for using the Euclidean distance as the distance metric. Comparing them to Figure 3, it can be seen that the cosine similarity and the Euclidean distance are consistently outperformed by the KL divergence distance metric when cluster seeds are given. Moreover, evaluation metrics on the cosine similarity and the Euclidean distance often have larger statistical variations (bigger confidence intervals) than those on the KL divergence. This further suggests that KL divergence as the distance metric tends to be more statistically stable on our dataset.

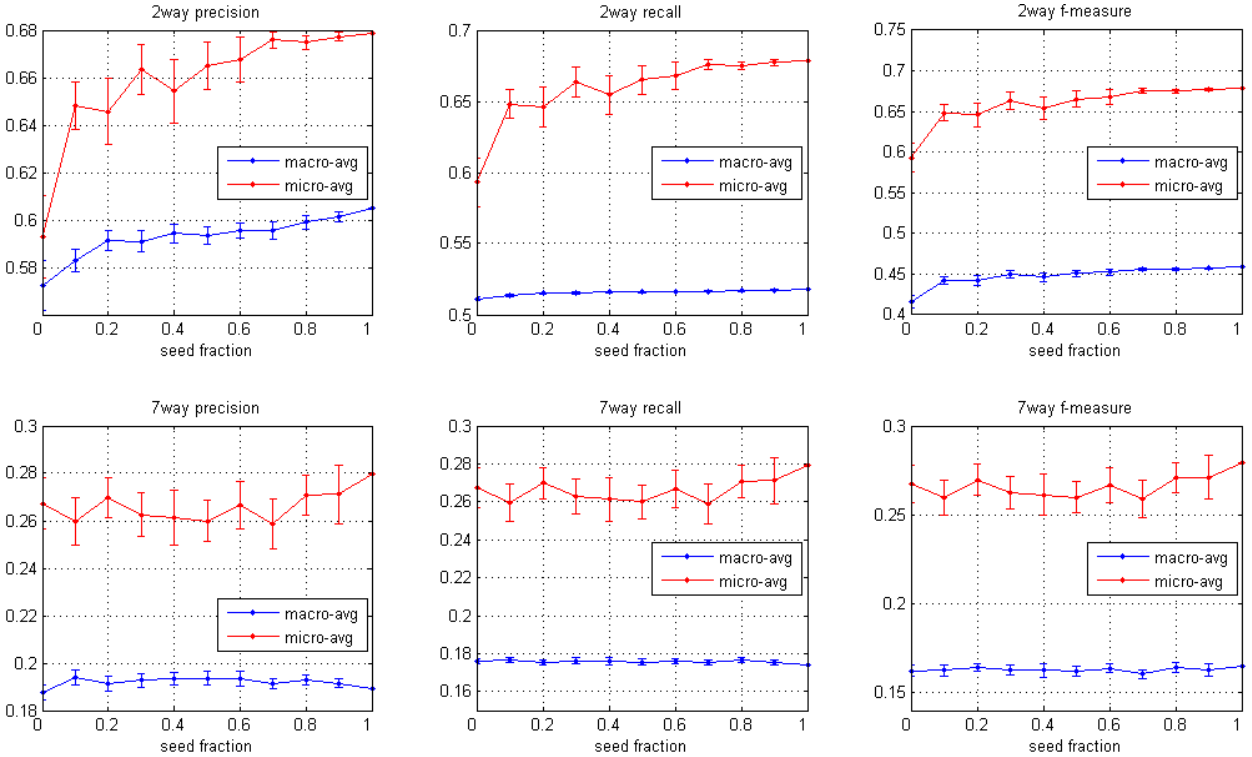


Figure 4 Macro-averaged and micro-averaged Precision, Recall and F-measure on 2-way and 7-way relation clustering using cosine similarity as distance. Results are averaged over 30 runs, confidence intervals at α=0.05 are also shown.

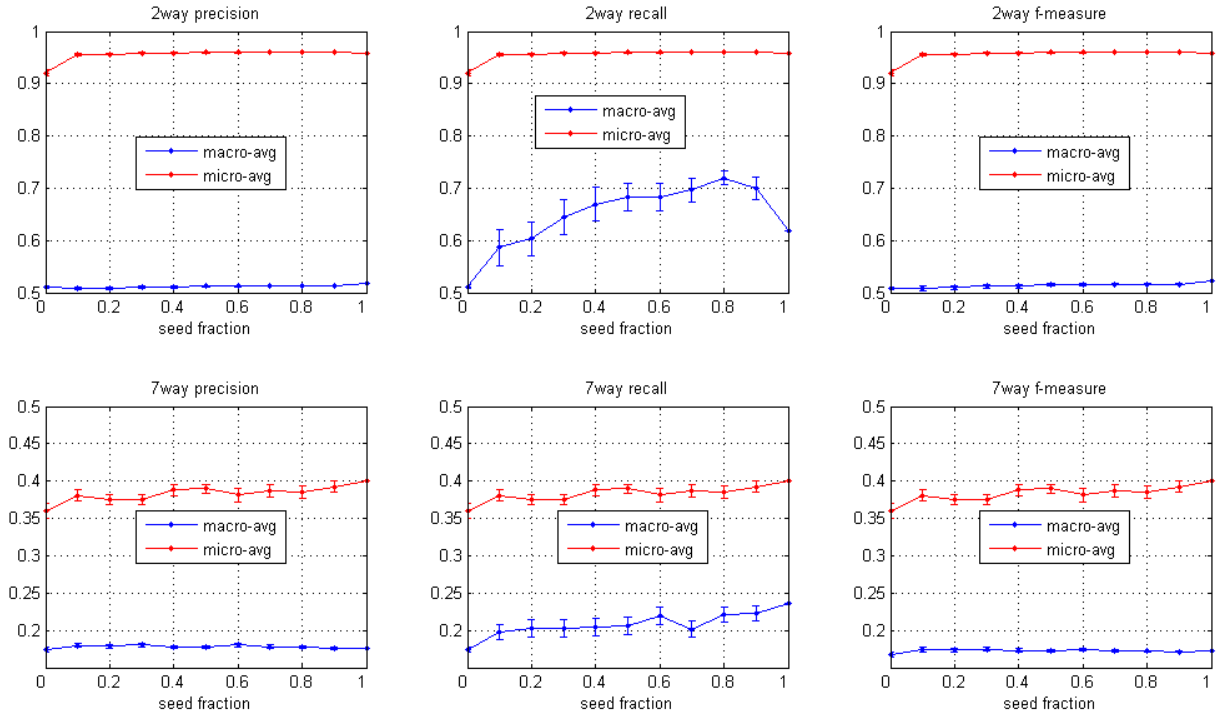


Figure 5 Macro-averaged and micro-averaged Precision, Recall and F-measure on 2-way and 7-way relation clustering using the Euclidean distance. Results are averaged over 30 runs, confidence intervals at α=0.05 are also shown.

Open Source Clinical NLP – More than Any Single System

James Masanz, MS[1], Serguei V. Pakhomov, PhD[2], Hua Xu, PhD[3], Stephen T. Wu, PhD[1], Christopher G. Chute, MD DrPH[1], Hongfang Liu, PhD[1]
[1]Mayo Clinic, Rochester, MN, [2]College of Pharmacy and Institute for Health Informatics, University of Minnesota, [3]School of Biomedical Informatics in The University of Texas Health Science Center at Houston

Abstract

The number of Natural Language Processing (NLP) tools and systems for processing clinical free-text has grown as interest and processing capability have surged. Unfortunately any two systems typically cannot simply interoperate, even when both are built upon a framework designed to facilitate the creation of pluggable components. We present two ongoing activities promoting open source clinical NLP. The Open Health Natural Language Processing (OHNLP) Consortium was originally founded to foster a collaborative community around clinical NLP, releasing UIMA-based open source software. OHNLP's mission currently includes maintaining a catalog of clinical NLP software and providing interfaces to simplify the interaction of NLP systems. Meanwhile, Apache cTAKES aims to integrate best-of-breed annotators, providing a world-class NLP system for accessing clinical information within free-text. These two activities are complementary. OHNLP promotes open source clinical NLP activities in the research community and Apache cTAKES bridges research to the health information technology (HIT) practice.

1. Introduction

Rapid growth in the clinical implementation of large electronic medical records (EMRs) has led to an unprecedented expansion in the availability of dense longitudinal datasets for clinical and translational research[1, 2]. This growth is being fueled by recent federal legislation that provides generous financial incentives to institutions demonstrating aggressive application and "meaningful use" of comprehensive EMRs[3-5]. Efforts are already underway to link these EMRs across institutions and to standardize the definition of phenotypes for large scale studies of disease onset and treatment outcome, specifically within the context of routine clinical care[6-8]. A well-known challenge in the secondary use of EMR data for clinical and translational research is that much of detailed patient information is embedded in narrative text. It is a very time-consuming and costly process to manually extract information from clinical records[9-11]. Researchers have used NLP systems to identify clinical syndromes and common biomedical concepts from radiology reports, discharge summaries, problem lists, nursing documentation, and medical education documents[12-18].

Different NLP systems have been developed at different institutions and used to convert clinical narrative text into structured data that may then be used for other clinical applications and studies. The systems include MedLEE[19, 20], MetaMap[21], KnowledgeMap[22], Apache cTAKES[23,24], and HiTEX[25]. Successful stories in applying NLP to clinical and translational research have been reported widely, ranging from identifying patient safety occurrences[26] to facilitating genomics research such as gene-disease association analysis and pharmacogenomic studies[15-18, 27, 28]. However, the lack of interoperability of NLP systems limits the full potential of applying NLP for clinical and translational research.

Multiple national-level informatics initiatives aim to enable the secondary use of EMRs for clinical and translational research but lack dedicated effort in addressing interoperability of existing NLP systems. The i2b2 (Informatics for Integrating Biology and the Bedside) center, one of the NIH roadmap centers, has developed a scalable informatics framework for clinical and translation research. One NLP system, HiTEX, has been distributed through the i2b2 platform as an optional component, and Apache cTAKES has become the platform of choice for i2b2. The NLP activities in iDASH (integrating Data for Analysis, Anonymization, and Sharing), another roadmap national center funded by NIH, have been focusing on developing an NLP ecosystem to i) develop and disseminate shareable NLP tools and annotated data, ii) determine what gaps exist between current efforts and an ideal software/data ecosystem, and iii) outline needs for integrating our efforts, data, and tools. Another informatics initiative, SHARP Area 4 project headed by Mayo Clinic has focused on building infrastructure tools including NLP through Apache cTAKES for high throughput phenotyping. The Consortium for Healthcare Informatics Research (CHIR) has been adopting the best-of-breed NLP methods to unlock clinical information available as free text in the VA EMR system. All those efforts promote uses of NLP for clinical investigation, point-of-care applications, and support for decision support[29,30].

We believe there are two factors causing the lack of interoperability among existing NLP systems: i) closed source development and ii) lack of standardization of NLP data models and exchange format. The recent development and adoption of open source engineering framework architectures such as GATE (General Architecture for Text Engineering)[31] and UIMA (Unstructured Information Management Architecture)[32] in the research community has enabled scalable and modular development of tools for processing unstructured information (text included). In the biomedical domain, several groups have successfully implemented NLP platforms based on UIMA or developing wrappers around existing open source NLP tools[33-35].

In this paper, we present two ongoing activities to promote open source interoperable clinical NLP research and development. One is the development of the infrastructure for The Open Health Natural Language Processing (OHNLP) Consortium and the other is the release of cTAKES as a top-level project within the Apache Software Foundation. In the following, we first provide some background information. We then present the ongoing activities in OHNLP and cTAKES in more detail.

2. Background and Motivation

The Open Health Natural Language Processing (OHNLP) Consortium was founded in 2009 to release open source clinical NLP tools developed under UIMA. Two UIMA-based NLP systems with distinct NLP data models were initially contributed to OHNLP. One is cTAKES (clinical Text Analysis and Knowledge Extraction System) and the other is MedKAT/p[36]. The two systems were not able to directly communicate with each other; even though both were built upon the UIMA framework. The reasons for this were:

- **Independent development with different purposes** - cTAKES and MedKAT/p were developed at different times and with different aims, even though there was some overlap in the people involved in both. cTAKES was initially developed to process clinical notes, while MedKAT/p was written to process pathology reports. This resulted in the data models have different focuses, and not being aligned.
- **Java naming conventions** - The two systems were written in Java but not owned by a single institution. Java naming conventions generally use institution names within the source code (using com.ibm and edu.mayo within the Java package names); the names used within the two data models (i.e., UIMA type systems) had no chance of being the same. For the two systems to share data seamlessly requires either changing the source code of one or both of them, or writing some interfacing code.

When all components for a UIMA pipeline come from a single development team, the team can agree on the type system, and data can flow seamlessly between components. When components are developed by different teams, often some interfacing code must be written to convert data from one structure or format to another. Apart from the data type disparities, interoperability between NLP systems may be affected by use of different linguistic theories and approaches to representing linguistic entities. For example, one NLP system may rely on the original Penn Treebank[37] tags for part-of-speech tagging, whereas another system may use a modification of Penn Treebank or an entirely different tag set altogether. This is a particularly problematic interoperability issue because the disparity between tag sets may not necessarily lead to an overt and easily detectable system failure, but instead result in unpredictable and possibly less accurate system performance. Therefore, raising the awareness in the NLP community of these issues is an important first step toward creating interoperable NLP systems and components. The interoperability must exist between NLP system designers and developers first in order for the interoperability between NLP systems to follow.

There are multiple activities underway working toward interoperable clinical NLP. Specifically, under the SHARPn project, an NLP common data model was developed to be more comprehensive while still allowing for extensions[38]. That data model has been adopted by Apache cTAKES as well as multiple open source NLP tools released under OHNLP including MedTagger, MedXN, and MedTime.

To create community awareness of existing informatics tools, the ORBIT site (hosted by National Library of Medicine) catalogs informatics software, knowledge bases, data sets and design resources[39]. However, the ORBIT site does not provide any hosting services for projects themselves. Developers of informatics software who wish to release the software open source must decide upon a code repository, a site for documentation, how to provide a place for questions to be asked, etc. The iDASH Natural Language Processing (NLP) Ecosystem provides a link to the ORBIT registry as well as a virtual machine image that you can download which contains several NLP tools[40]. It intends to also host source code; however it does not yet host source code, nor does it provide a way to download the tools individually.

The Apache Software Foundation (ASF) - The ASF provides Wiki space, mailing lists, options for issue tracking (JIRA and Bugzilla), worldwide mirroring of distribution website, press releases for new projects, and expertise in a variety of areas such as licensing and trademarks. The ASF's mission is to provide software for the public good. The ASF however requires its projects have a diverse set of committers. The ASF's guide to creating a proposal for a new Apache project states "Apache is interested only in communities. Candidates should start with a community and have the potential to grow and renew this community by attracting new users and developers."(41) Tools are typically not shared through ASF until a certain level of maturity has been reached.

3. Towards Community-Driven Open Source Clinical NLP

Figure 1 illustrates the history as well as our thought towards open source clinical NLP which was pioneered by the establishment of OHNLP. Initially OHNLP was home to two clinical NLP systems – cTAKES and MedKAT/p. Other systems have since been contributed to OHNLP, and cTAKES has moved on to the ASF and become Apache cTAKES. OHNLP has added to its original mission – it will be not only a source of information for clinical NLP software outside of OHNLP, but it will also be the source of interfaces that address interoperability between existing NLP software. Therefore, OHNLP can be viewed not only as a home for mature medical NLP tools but also as an innovation incubator for projects in early experimental stages of development

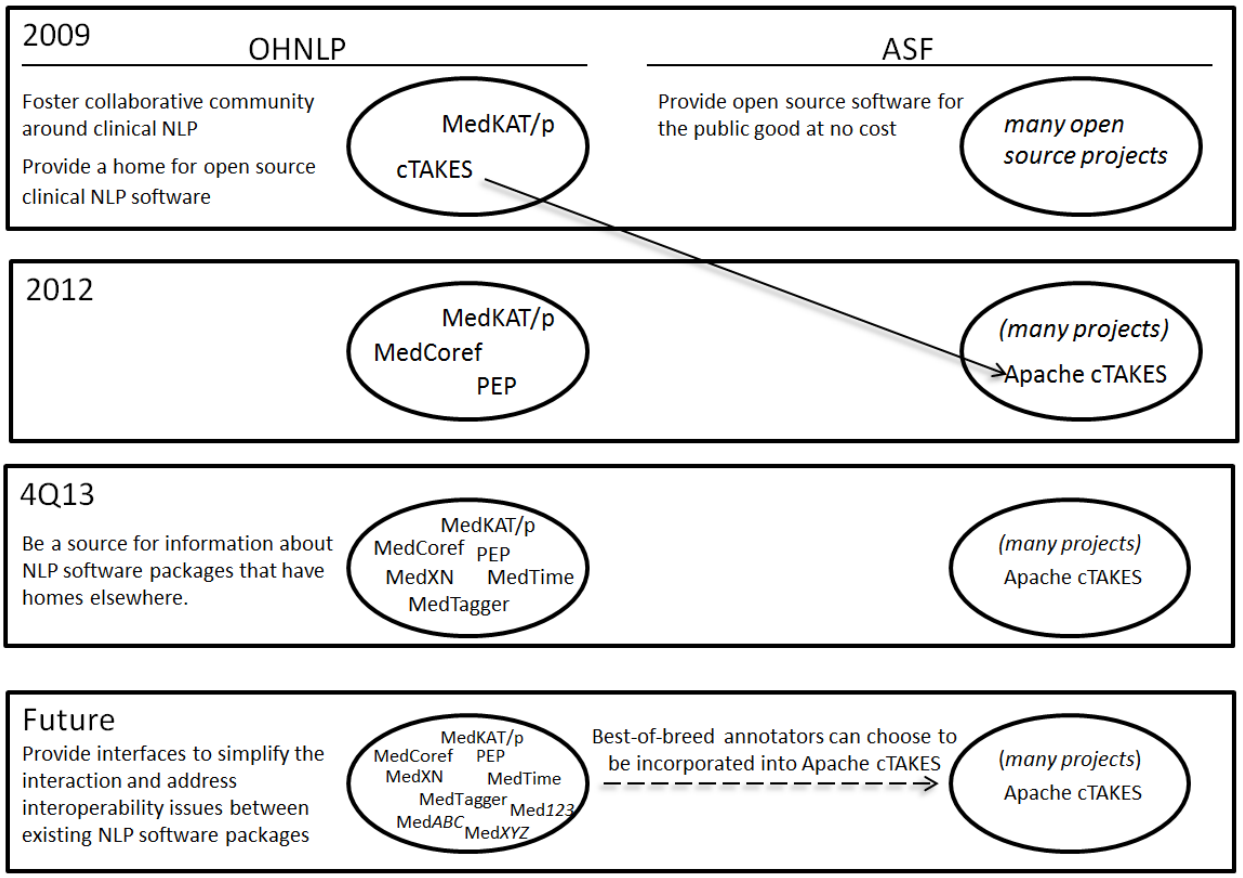


Figure 1. OHNLP and Apache cTAKES Timeline

3.1. Apache cTAKES

After an initial innovation incubation phase at OHNLP, cTAKES has migrated to Apache cTAKES, which strives to be a world-class clinical Natural Language Processing (NLP) system, containing best-of-breed annotators. It is open source and the Apache cTAKES community welcomes contributions. Apache cTAKES is modular and expandable at the information model and method level. As an Apache project, it is licensed under the Apache License, Version 2.0 by the Apache Software Foundation (ASF), which is how the ASF licenses all its current projects .

Apache cTAKES aims to be modular and expandable at the information model and method level and can be used in a great variety of retrievals and use cases. The Apache cTAKES community is committed to best practices and R&D (research and development) by using cutting edge technologies and novel research and facilitating the translation of the best performing methods into the project.

Apache cTAKES has incorporated and intends to continue to incorporate best-of-breed annotators. The community is multi-institutional or non-institutional – not dependent upon any single institution for funding. One goal is to align with industry/community standards and conventions.

Coinciding with cTAKES becoming an Apache project, work was done to generalize the type system (i.e., data model). However there are still many different type systems in use by different tools and systems. Work to have a basic common type system across several systems is under discussion – it is only in the planning stage and only addresses a few systems.

3.2. The OHNLP Consortium

The mission of the current OHNLP Consortium consists of three main objectives:

1. Provide a home for existing and emerging clinical NLP software.

2. Provide a gateway to NLP software packages that reside in other repositories.

3. Provide interfaces to simplify the interaction of various NLP packages and address interoperability issues between existing NLP software packages; allowing seamless information exchange among them.

Be a home for open source clinical NLP software. Existing open source repositories such as the Sourceforge provide open source software developers a place to host source code, publish releases, maintain documentation, and track issues. Creating a new Sourceforge project does not provide the visibility that comes with being associated with the OHNLP Consortium. For those tools/systems looking for a home, OHNLP can be the place for the tool/system to live, and for a community to develop around it. OHNLP has a dedicated website with a Wiki, and a project (https://sourceforge.net/projects/ohnlp/) at Sourceforge with:

- a repository for the code (an SVN instance)
- issue (bug) trackers
- forums
- release staging and publishing

While the OHNLP Consortium encourages the use of the Apache License, OHNLP does not comprise a single system, therefore the consortium does not prohibit individual systems or tools that are contributed to it from being licensed under the GNU General Public License (GPL) or other licenses that would be problematic for Apache. OHNLP can provide a home for these tools/systems. OHNLP can be the long term home of an NLP tool/system, or it can be a starting place of a tool/system that moves on to somewhere else such as the ASF. There is no need for a tool's owner to decide ahead of time.

Be a registry for existing clinical NLP systems, tools, and resources. OHNLP is intended as a hub for clinical NLP resources and tools; aiming to facilitate sharing of resources and of software. And beyond simply sharing software packages (binaries), one of OHNLP's goals is to encourage the developers of clinical NLP software to release their code as open source. If an institution requires others to sign a data use agreement (DUA) before accessing a shared resource, OHNLP can be the home for the DUA request forms and contact information.

Promote interoperability by a common data model and address interoperability among existing NLP systems. There is a need for existing NLP tools/systems to be able to share data (interoperate) with each other, including Apache cTAKES. But there is much work to be done to make existing clinical NLP tools interoperate better. Once tools and systems are established it is often difficult to find the time to work on interoperability. A common data model (OHNLP types) built upon the SHARPn common data model is currently available at OHNLP. Its intent is not to dictate the use of specific data types and structures but to provide a starting point and make it easier for developers to consider interoperability with other systems as one of their objectives. Thus, developers can extend the common data model for their specific systems or build their systems in a way that will not preclude future integration with other tools in the OHNLP initiative. There are two options for improving how existing systems interoperate – develop a common type system/data structure and rework the existing systems to use them, or develop

interfaces to allow the existing systems to interoperate. Developers should be able to exercise either of the options for improving the interoperability. For example, the NLP team at Mayo Clinic recently reworked their NLP systems including MedTagger, MedTime, and MedXN to adopt the common data model. Currently, OHNLP is in the process of developing interfaces to multiple open source clinical NLP systems including Apache cTAKES, MedEx, BioMedICUS, MedLEE, MetaMap, and HiTEX.

4. Conclusion

We have described two ongoing activities for open source clinical NLP. The OHNLP Consortium provides to the clinical NLP community resources that are not found elsewhere under a single umbrella. OHNLP provides visibility to NLP software and OHNLP infrastructure includes a Wiki, a registry of related software, and a place where a tool or system without an established diverse development community – including tools developed by a single author – can be hosted. Meanwhile, Apache cTAKES aims to provide best-of-breed NLP modules to the community and facilitates the translation of research into practice.

Future work would include the continuous development of both OHNLP and Apache cTAKES and the collaboration with other initiatives to advance open source clinical NLP.

5. Acknowledgements

We would like to thank and acknowledge the work of Anni Coden, Ph.D. Michael Tanenblatt, Igor Sominsky, et al. at IBM as well as Guergana Savova, Ph.D. and others involved in founding the OHNLP Consortium. We would also like to thank and acknowledge the work of Guergana Savova and Pei Chen of Boston Children's Hospital and Harvard Medical School as well as others in transitioning cTAKES to Apache cTAKES. The work on OHNLP is funded in part by R01 GM102282. The work on Apache cTAKES was funded in part by the SHARPn (Strategic Health IT Advanced Research Projects) Area 4: Secondary Use of EHR Data Cooperative Agreement from the HHS Office of the National Coordinator, Washington, DC. DHHS 90TR000201.

References

1. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. New England Journal of Medicine. 2010;362(3):192-5.
2. Jha AK, DesRoches CM, Campbell EG, et al. Use of electronic health records in US hospitals. New England Journal of Medicine. 2009;360(16):1628-38.
3. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. N Engl J Med. Jan 21;362(3):192-5.
4. Secretary Sebelius Announces Final Rules To Support 'Meaningful Use' of Electronic Health Records. US Department of Health and Human Service 2010 [cited 2010 10/15]; Available from: http://www.hhs.gov/news/press/2010pres/07/20100713a.html
5. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. Journal of the American Medical Informatics Association : JAMIA. 2010 Mar-Apr;17(2):131-5.
6. McCarty CA, Wilke RA. Biobanking and pharmacogenomics. Pharmacogenomics. 2010;11(5):637-41.
7. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. The American Journal of Human Genetics. 2010;86(4):560-72.
8. Pace WD, Cifuentes M, Valuck RJ, Staton EW, Brandt EC, West DR. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009;151(5):338.
9. South BR, Shen S, Jones M, et al. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics. 2009;10(Suppl 9):S12.
10. Grishman R, Huttunen S, Yangarber R. Information extraction for enhanced access to disease outbreak reports. J Biomed Inform. 2002;35(4):236-46.
11. Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. Journal of the American Medical Informatics Association. 2011;18(4):387-91.
12. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000:270-4.
13. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
14. Chapman W, Bridewell W, Hanbury P, Cooper G, Buchanan B. A simple algorithm for identifying negated findings and diseases in discharge summaries. Journal of biomedical informatics. 2001;34(5):301-10.

15. Ananthakrishnan AN, Cai T, Savova G, et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: a novel informatics approach. Inflammatory bowel diseases. 2013;19(7):1411-20.
16. Ananthakrishnan A, Gainer V, Perez R, et al. Psychiatric co-morbidity is associated with increased risk of surgery in Crohn's disease. Alimentary pharmacology & therapeutics. 2013;37(4):445-54.
17. Savova GK, Olson JE, Murphy SP, et al. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. Journal of the American Medical Informatics Association. 2012;19(e1):e83-e9.
18. Lin C, Karlson EW, Canhao H, et al. Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. PLoS One. 2013;8(8):e69932.
19. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods Inf Med. 1998 Nov;37(4-5):334-44.
20. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:270-4.
21. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. May 1;17(3):229-36.
22. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A, 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2003:195-9.
23. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):507-13.
24. Apache cTAKES. http://ctakes.apache.org/. 2013
25. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006:931.
26. Murff HJ, FitzHenry F, Matheny ME, et al. Automated Identification of Postoperative Complications Within an Electronic Medical Record Using Natural Language Processing. JAMA: The Journal of the American Medical Association. 2011;306(8):848-55.
27. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;26(9):1205.
28. Xu H, Jiang M, Oetjens M, et al. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. Journal of the American Medical Informatics Association : JAMIA. 2011 Jul-Aug;18(4):387-91.
29. Wagholikar KB, MacLaughlin KL, Henry MR, et al. Clinical decision support with automated text processing for cervical cancer screening. Journal of the American Medical Informatics Association. 2012;19(5):833-9.
30. Vleck TV, Elhadad N. Corpus-based problem selection for EHR note summarization. In: AMIA annu symp proc; 2010. p. 817–21.
31. Cunningham DH, Maynard DD, Bontcheva DK, Tablan MV. GATE: A framework and graphical development environment for robust NLP tools and applications. 2002.
32. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering. 2004;10(3-4):327-48.
33. Kano Y, Baumgartner WA, McCrohon L, et al. U-Compare: share and compare text mining tools with UIMA. Bioinformatics. 2009;25(15):1997.
34. Hahn U, Buyko E, Landefeld R, et al. An overview of JCoRe, the JULIE lab UIMA component repository. 2008; 2008. p. 1–7.
35. v3NLP. https://wiki.chpc.utah.edu/display/htx/v3NLP+Framework+Tool+Development (accessed 1 Oct 2013). 2013
36. Coden A, Savova G, Sominsky I, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform. 2009;42(5):937-49.
37. Penn. http://repository.upenn.edu/cis_reports/570/. 2013
38. Wu ST, Kaggal VC, Dligach D, et al. A common type system for clinical natural language processing. J Biomed Semantics. 2013;4(1):1.
39. ORBIT. National Library of Medicine. ORBIT: Online Registry of Biomedical Informatics Tools. 2011. http://orbit.nlm.nih.gov 2013

40. NLP Ecosystem. http://nlp-ecosystem.ucsd.edu/events/webinars/making-natural-language-processing-nlp-more-accessible-analysis-clinical-text. 2012
41. Apache. http://incubator.apache.org/guides/proposal.html. 2013

Towards Transforming Expert-based Content to Evidence-based Content

Soheil Moosavinasab, BS[1, 2]; Majid Rastegar-Mojarad, MS[1, 2]; Hongfang Liu, PhD[1]; Siddhartha R. Jonnalagadda, PhD[1, 3]

[1] Department of Health Sciences Research, Mayo Clinic, Rochester, MN

[2] University of Wisconsin-Milwaukee, Milwaukee, WI

[3] Northwestern University Feinberg School of Medicine, Chicago, IL

Abstract

The goal of this paper is to find relevant citations for clinicians' written content and make it more reliable by adding scientific articles as references and enabling the clinicians to easily update it using new information. The proposed approach uses information retrieval and ranking techniques to extract and rank relevant citations from MEDLINE for any given sentence. Additionally, this system extracts snippets of relevant content from ranked citations. We assessed our approach on 4,697 MEDLINE papers and their corresponding full-text on the subject of Heart Failure. We implemented multi-level and weight ranking algorithms to rank the citations. We demonstrate that using journal relevance and study design type improves results obtained from only using content similarity by approximately 40%. We also show that using full-text, rather than abstract text, leads to extracting higher quality snippets.

Introduction

In this paper, we developed a system, known as CiteFinder, to find citations for clinical sentences. For each given sentence, the system finds citations from MEDLINE articles, ranks the citations based on similarity with the sentence, and extracts a snippet for each citation. We implemented a tool for the system that allows the user to submit a sentence and receive back the top relevant citations. This aids in transforming the expert-based content (a paradigm not used by certain clinical knowledge systems such as UpToDate©[1], but relatively common among some care providers[2]) to evidence-based content – the accepted paradigm[3]. This will offer clinicians the flexibility of easily authoring evidence-based guidance and FAQs for their peers.

Background

Citation finding has been investigated to recommend relevant papers to researchers[4–8]. There are also studies on information retrieval in the medical domain. For example, Plaza and Diaz[9] proposed a method to query similar Electronic Health Records using UMLS concepts. Hersh and Hickam[10] studied the effectiveness of electronic information retrieval systems for physicians. Lu[11] investigated web tools for searches in biomedical literature. Bachmann et al [12] proposed and validated search strategies used to identify diagnostic articles recorded on MEDLINE, with special emphasis on precision. Bernstam et al[13] studied how citation-based algorithms that are developed to extract relevant and important citations for the World Wide Web are useful in the biomedical literature domain. They compared eight citation algorithms, including simple PubMed queries, clinical queries, citation counts, journal impact factors, etc. Their research concluded that these citation-based algorithms are useful in the domain of biomedical literature. Lin et al[14] extracted relevant MEDLINE citations and ranked them based on several ranking methods, including citation counts per year and journal impact factors. Darmoni et al[15] used MeSH concepts for indexing and information retrieval. Some studies have also been conducted on query expansion using MeSH terms in PubMed. Lu et al[16] analyzed the effect of using MeSH terms in a PubMed automatic search. In the current study, we also used MeSH concepts to find relevant citations.

Methods

CiteFinder consists of four major parts: sentence expansion, citation extraction, citation ranking, and snippet generation.

After a user submits a sentence (although technically this could be applied for an entire paragraph), the system finds relevant citations for the sentence from our collection of MEDLINE articles. To find relevant citations, MeSH terms are used. CiteFinder extracts MeSH terms from the sentence and searches them in MeSH terms of each indexed MEDLINE article. Then it ranks the articles based on three measures: MeSH terms, journal relevance, and epidemiological study design[17]. The final step is producing snippets for the retrieved citations based on the extracted major terms (mentions) of the sentence.

Figure 1 illustrates the architecture of the system. We use a running example in Appendix 1 to clarify each part of the system.

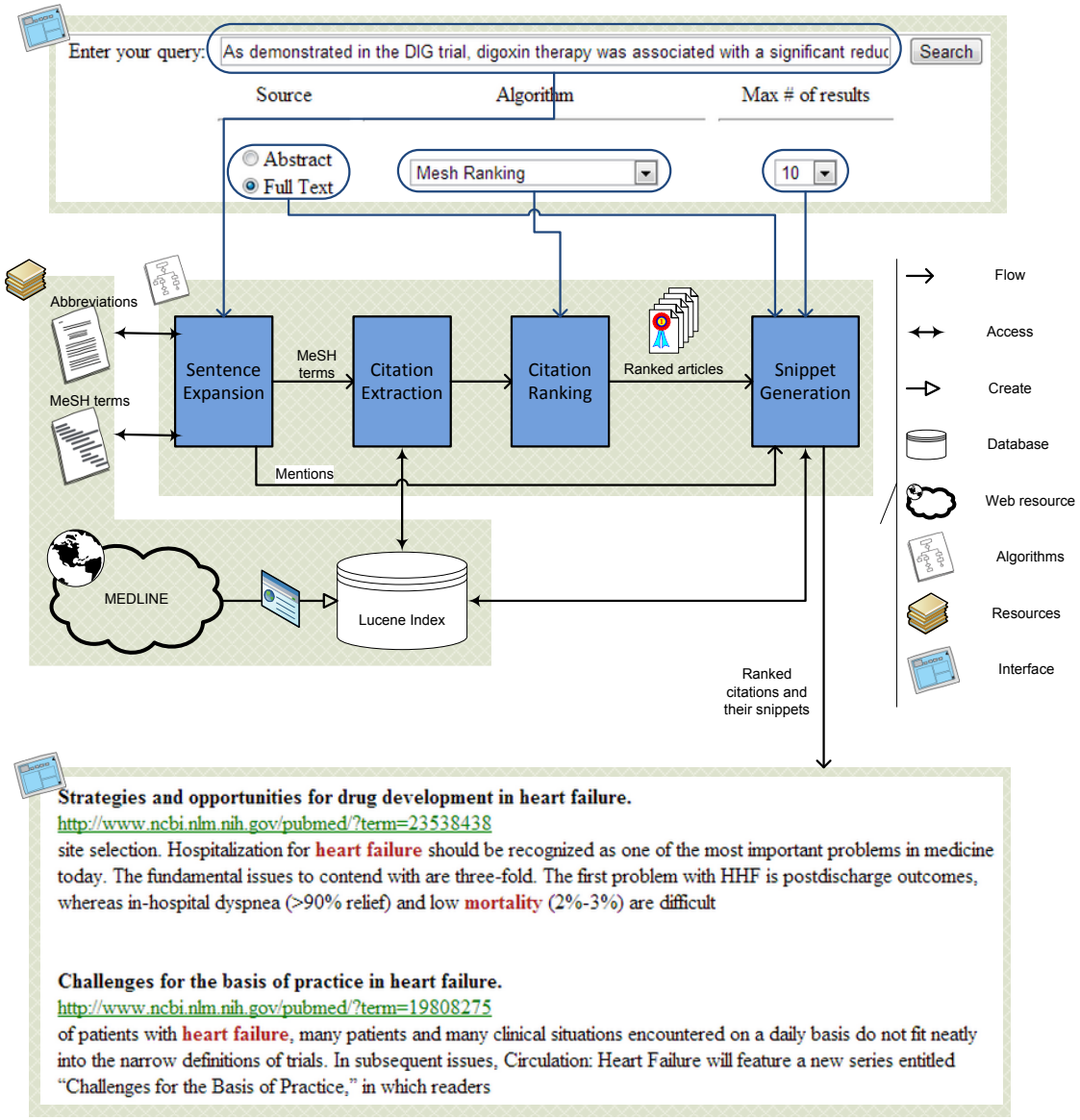


Figure 1: System Architecture The figure illustrates the sentence expansion, citation extraction, citation ranking, and snippet generation components and their integration with the user-interface – all of them available at http://sourceforge.net/projects/cksauthorer.

Step 1: Sentence Expansion Since each word in a sentence might not be in an article or abstract, we locate important terms, normalize them and expand. That is, the sentence goes through OpenNLP tokenization[18], lexical normalization[19], dictionary-based concept extraction using both UMLS Metathesaurus and MeSH using Aho-Corasick algorithm[20], and abbreviation expansion (using a list of 6,024 abbreviations and their full-forms derived from UMLS).

Step 2: Citation Extraction The next step is to find relevant citations for the sentence based on the extracted MeSH terms. To be able to generalize the system to other documents such as textbooks and guidelines and build a fast system, we indexed MEDLINE abstracts and their full-text with Lucene[21]. CiteFinder stores the text, title,

publication type, and MeSH terms of each article. The articles with at least one MeSH term in common with the sentence will be retrieved at this step.

Step 3: Citation Ranking In order to rank the retrieved citations with regard to their importance and similarity with the sentence, three measures are applied: MeSH ranking, journal relevance, and study design. In the following section, we describe each of them and explain how we calculate a score.

Measure 1: MeSH Measure. The MeSH measure shows the semantic similarity of the sentence and articles. We use the score calculated by Lucene for each returned article from the MeSH extraction step. Our language model that is based on Mesh terms as opposed to individual words built from Lucene[22] takes into account both the TF-IDF (frequency of the term in the document with penalty to each term if it is commonly occurring in other documents), and Number of MeSH terms in an article. This performed better than TF-IDF over individual words.

Measure 2: Journal Relevance. The idea behind this measure is that a citation that is published in a high-quality journal has extra chance to obtain a higher rank than a citation with the same MeSH score published in a low-quality journal in specific domain (example, Heart Failure [CHF]). We previously studied the task of prioritizing journals and obtained a formula to rank each journal[23] based on information available from Scopus[24] and PubMed – Journal Relevance score = (0.82640 * SCImago Journal Ranking) – (0.00377 * Number of articles) + (0.00258 * Number of articles for 3 years) – (0.00190 * Number of cited-articles for 3 years) – (0.01846 * Number of references per article) + (0.00295 * Number of CHF-indexed Medline abstracts) + (0.62864 * Is Broad Journal Heading cardiology?) – (0.32753 * Is Core clinical journal?).

Measure 3: Study Design. It is well known that the strength of the findings in clinical research depends on the study design and follows this order: systematic review, randomized controlled trial, multiple time series, nonrandomized trial, cohort, case-control, time series, cross-sectional, and case series[17]. Weight levels 9 to 1 are assigned to each study type, respectively. To decide on the study type of a citation, we consider several sections of articles, including publication type, abstract text, MeSH headings, and article title. A publication for which no study design is detected gets the least possible score of 1.

Ranking Methods

We proposed two ranking schemes using the above measures to assign ranks to retrieved citations. It should be noted that all scores of the measures are normalized to the range of 0 to 1.

1) ***Multi-Level Ranking.*** A multi-level approach ranks the articles in a cascade trend. The idea is to rank the articles with one measure, and then split the sorted articles into brackets and re-rank the brackets with scores obtained from other measures. Finding the best bracket size for each level is one of the challenges of this approach. In this experiment, after extracting and ranking citations via MeSH measure, CiteFinder splits them in N brackets based on their MeSH score. Table 1 shows the results with different variation of N. In the next step, the journal measure is used to rank the citations within the bracket. In the last step, the study design measure is used to rank the citations in each newly created N brackets to produce the final list of ranked citations.
2) ***Weight Ranking.*** In the second approach, the final score is calculated using the formula: Score = (MeSH weight * MeSH score) + (Journal Relevance weight * Journal Relevance score) + (Study Design weight * Study Design score). This approach is valid considering that these three metrics are independent and orthogonal.

Step 4: Snippet Generation. Snippet generation is helpful for clinicians to get an idea about the existence of similar information in scholarly articles and improve their written content. A query made by disjunction of the extracted mentions is used to extract a maximum of three snippets for each article. See Appendix 1 for a running example of all the steps.

Evaluation

Data Collection. CiteFinder contains 4,697 MEDLINE papers about Heart Failure. This corpus includes two major sources (the duplicated articles or the articles with only scanned-version availability have been removed):

- 2,582 articles retrieved by “heart failure[MeSH Major Topic]” query at PubMed Central

- 2,262 articles retrieved by “Congestive Heart Failure[MeSH Major Topic]” query at PubMed on four top ranked journals for CHF topic: 1. Circulation, 2. Circulation. Heart failure, 3. JAMA the Journal of the American Medical Association, and 4. The New England Journal of Medicine

Both the abstract and full-text of these papers are indexed separately with Lucene to allow us to compare the performance on both the abstract and full-text in extracting snippets.

Gold standard

The gold standard data contains 377 sentences referring to 456 citations. We primarily selected 7,864 sentences referring to 11,778 citations using all 150 retrieved articles from UpToDate© for the query – "heart failure". We then filtered out sentences with less than 15 words, less than 5 MeSH terms, or no full-text availability in our index files.

Results

To evaluate ranking methods, we consider median rank of expected citations for each sentence in our gold standard. If the expected citation of a sentence is not retrieved, its rank is assumed to be the worst (lowest). So we consider the median rank of all citations in the gold standard, regardless of whether the system finds and ranks them or not. In this scenario, we were unable to find 5.26% (24 of 456) of the citations, but the currently reported median ranking is affected by recall.

Multi-Level Ranking

In the first experiment, we explored multi-level ranking. Table 1 shows median rank for the multi-level ranking approach. Both Journal Relevance and Study Design show improvement in the results.

Table 1: Multi-level ranking results

Measures \ # of Brackets	10	20	100
MeSH	76	76	76
MeSH and Journal Relevance	66	67	65
MeSH and Study Design	73	71	67
MeSH, Journal Relevance, and Study Design	**64**	**65**	**65**

Weight Ranking

In the second experiment, we attempted to find the best coefficient for Journal Relevance when the MeSH coefficient is 1. A range of coefficients between 0 and 2 were explored, and the results indicated 0.5 as the best weight for Journal Relevance Figure 2 illustrates these results.

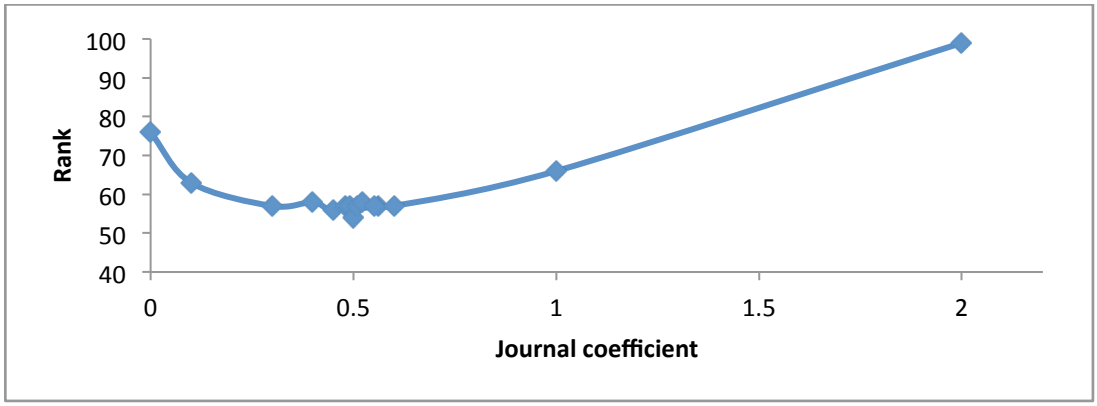


Figure 2: The chart indicates how different journals coefficients affect median ranking in our gold standard. In this experiment, the coefficient for MeSH measure is 1.

Then we used the best combination (MeSH=1, Journal Relevance=0.5) as the constant and found the best weight for Study Design (0.30). Figure 3 indicates the results of this experiment.

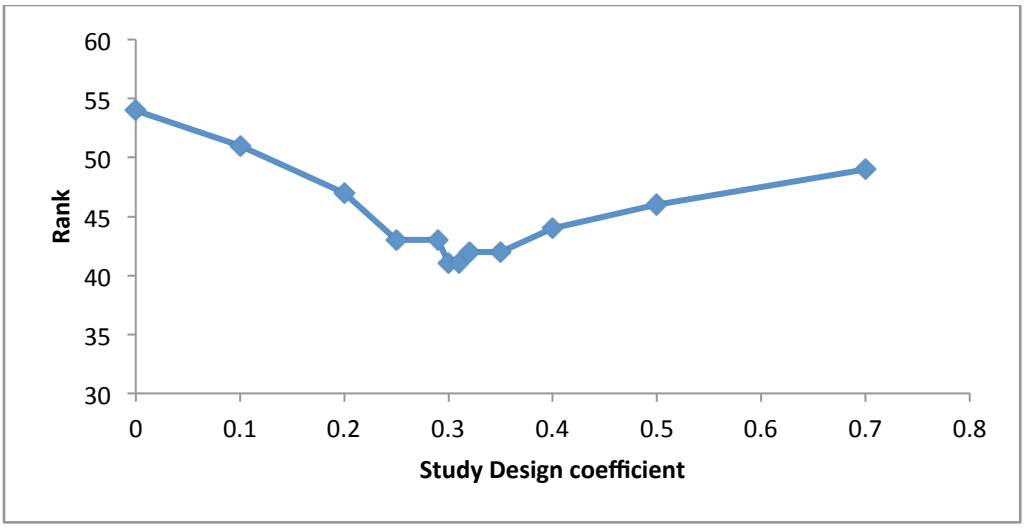


Figure 3: The chart indicates how different Study Design coefficients affect median ranking in our gold standard. In this experiment, the coefficient for MeSH measure is 1 and Journal Relevance is 0.5.

Snippet Generation After ranking returned citations, we extracted snippets for each of them. In this experiment, we explored whether using the full-text for extracting snippets is better than using the abstract. The experiment on the gold standard indicated that when CiteFinder uses full-text, it is able to extract at least one snippet for 99.7% of citations (in 431 of 432 extracted and ranked citations). When the system looks for snippets in an abstract, it extracted snippets for 80.7% of the citations (349 out of 432). Further, as the system tries to extract the best snippet (as adjudicated by the MeSH-based Lucene similarity), we discovered that only 22.58% of the best snippets come from the abstract text with the rest coming from full-text. This means that using the full-text instead of abstract text leads to the collection of more and better snippets.

Discussion

Ranking Algorithm. We implemented both multi-level and weight ranking algorithms to rank the citations. Results show more improvement in the weight-ranking algorithm because of the flexibility of this approach to change the effectiveness of measures. On the other hand, the multi-level approach is sensitive to the number of results retrieved by CiteFinder. In cases where the number of retrieved articles is not considerably larger than number of brackets, the system will not actually utilize the second- or third-level measures.

Generalizability. The proposed system (CiteFinder) explores methods to find citations for sentences in the Heart Failure domain. Further experiments will be required to check the generalizability of the system in other domains. Future work should also explore better methods to infer the epidemiological study design of the publication and consider alternative ways to score them. Appendix 2 discusses further limitations.

Conclusions

Finding supporting citations for clinical sentences is challenging for clinicians. We propose a system (CiteFinder), which, after expanding a user's sentence, extracts relevant citations and ranks them to retrieve the best citation for a given sentence. This study demonstrates that using Journal Relevance and Study Design type will improve the MeSH term results by about 40% (from 76 to 41). We also show that using full-text instead of abstract-text helps in extracting better snippets; i.e., they have more pertinent information corresponding to the input queries. The code for various components including the user-interface is available at http://sourceforge.net/projects/cksauthorer.

Acknowledgments

This work was made possible by joint funding from the National Library of Medicine K99/R00 LM011389, National Institute of Health R01LM009959A1, and National Science Foundation ABI:0845523. We are also thankful to UpToDate© for allowing us to use their data for research purposes.

References

1. UpToDate [Internet]. Available from: http://www.uptodate.com/home

2. Yeo GSH, Lim ML. Maternal and fetal best interests in day-to-day obstetrics. Ann Acad Med Singap 2011;40(1):43–9.

3. Lau J. Evidence-based medicine and meta-analysis: getting more out of the literature. Clinical decision support: the road ahead 2007;

4. Huang Z, Chung W, Ong T, Chen H. A graph-based recommender system for digital library. In: In Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press; 2002. p. 65–73.

5. Chen Y-L, Wei J-J, Wu S-Y, Hu Y-H. A similarity-based method for retrieving documents from the SCI/SSCI database. Journal of Information Science 2006;32(5):449–64.

6. Ratprasartporn N, Po J, Cakmak A, Bani-Ahmad S, Ozsoyoglu G. Context-based literature digital collection search. The VLDB Journal 2009;18(1):277–301.

7. Bollacker KD, Lawrence S, Giles CL. Discovering Relevant Scientific Literature on the Web. IEEE Intelligent Systems 2000;15(2):42–7.

8. Liang Y, Li Q, Qian T. Finding Relevant Papers Based on Citation Relations [Internet]. In: Wang H, Li S, Oyama S, Hu X, Qian T, editors. Web-Age Information Management. Springer Berlin Heidelberg; 2011 [cited 2013 Aug 14]. p. 403–14.Available from: http://link.springer.com/chapter/10.1007/978-3-642-23535-1_35

9. Plaza L, Díaz A. Retrieval of Similar Electronic Health Records Using UMLS Concept Graphs [Internet]. In: Hopfe CJ, Rezgui Y, Métais E, Preece A, Li H, editors. Natural Language Processing and Information Systems. Springer Berlin Heidelberg; 2010 [cited 2013 Aug 14]. p. 296–303.Available from: http://link.springer.com/chapter/10.1007/978-3-642-13881-2_31

10. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and systematic review. JAMA 1998;280(15):1347–52.

11. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford) 2011;2011:baq036.

12. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying diagnostic studies in MEDLINE: reducing the number needed to read. J Am Med Inform Assoc 2002;9(6):653–8.

13. Bernstam EV, Herskovic JR, Aphinyanaphongs Y, Aliferis CF, Sriram MG, Hersh WR. Using citation data to improve retrieval from MEDLINE. J Am Med Inform Assoc 2006;13(1):96–105.

14. Lin Y, Li W, Chen K, Liu Y. A Document Clustering and Ranking System for Exploring MEDLINE Citations. J Am Med Inform Assoc 2007;14(5):651–61.

15. Darmoni SJ, Soualmia LF, Letord C, et al. Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. J Med Libr Assoc 2012;100(3):176–83.

16. Lu Z, Kim W, Wilbur WJ. Evaluation of Query Expansion Using MeSH in PubMed. Inf Retr Boston 2009;12(1):69–80.

17. Fletcher RH, Fletcher SW, Fletcher GS. Clinical Epidemiology: The Essentials. Lippincott Williams & Wilkins; 2012.

18. Baldridge J, Morton T. OpenNLP. 2004.

19. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507–13.

20. Aho AV, Corasick MJ. Efficient string matching: an aid to bibliographic search. Commun ACM 1975;18(6):333–40.

21. Lucene [Internet]. Available from: http://lucene.apache.org/

22. Similarity in Lucene [Internet]. Available from: http://lucene.apache.org/core/3_0_3/api/core/org/apache/lucene/search/Similarity.html

23. Jonnalagadda S, Moosavinasab S, Li D, Abel M, Chute C, Liu H. Prioritizing journals relevant to a topic for addressing clinicians' information needs. 2013 (Under review).

24. SCImago. 2007;Available from: Available: http://www.scimagojr.com/

Appendix 1: Running example.

Here we use a running example to demonstrate input, output and results of the system in different steps.

Input query:

For patients who are still hypertensive after initiation of beta blockers and ACE inhibitors and/or ARBs or who cannot tolerate these drugs appropriate agents include loop diuretics nitrates hydralazine and some vasoselective calcium channel blockers (eg amlodipine and felodipine)

Expected citation for this sentence: *PMID9264493*

Abbreviations found:

ACE: angiotensin-converting enzyme

Extracted Mentions:

Hypertensive, initiation, beta blockers, ACE inhibitors, drugs, agents, loop diuretics, nitrates, hydralazine, calcium channel blockers, amlodipine, felodipine, angiotensin receptor blocker, angiotensin, receptor, blocker

Extracted MeSH Terms:

Adrenergic beta-Antagonists, Angiotensin-Converting Enzyme Inhibitors, Pharmaceutical Preparations, Diuretics, Nitrates, Hydralazine, Calcium Channel Blockers, Amlodipine, Felodipine

Rank using the multi-level ranking method: *6th*

Rank using the weight ranking method: *6th*

Query to extract snippet:

""CA" "blocker" "receptor" "angiotensin" "angiotensin receptor blocker" "felodipine" "amlodipine" "calcium channel blockers" "hydralazine" "nitrates" "loop diuretics" "agents" "drugs" "ACE inhibitors" "beta blockers" "initiation" "hypertensive""

The extracted snippet:

antagonists; use of blockers, long-acting nitrates, or other vasodilators (except ACE inhibitors...V-HeFT III Abstract Background Despite therapy with diuretics, ACE inhibitors and digoxin morbidity... or volume, which are reduced by nitrates and ACE inhibitors. Progressive LV remodeling is characterized

Appendix 2: Limitation of Measures.

MeSH Accessibility

MeSH measure is the main method we are using to rank the citations. Journal rank and study design type are added as a component to the MeSH measure to improve the results. All the articles that we have in our corpus are extracted from PubMed or PubMed Central provides MeSH terms for them. CiteFinder's main limitation is that if we want to expand the corpus to cover more articles from mentioned sources, we will need to use a MeSH extractor program to pull out and index the MeSH terms from the articles.

Journal Relevance Measure

We studied 23 sentences related to heart failure with 31 citations. The study shows that 31% of retrieved articles (12,362 of 39,839) were not from the 63 journals we already have. Having a list of important Heart Failure–related journals will automatically guarantee that many unavailable journals are not related to the query. Even though we should assign a score of zero to them, having a complete list of journals can improve the system.

Study Design

We assigned weights of 1 through 9 to different study design types. Machine Learning algorithms can be applied to assign more accurate and meaningful weights to the elements.

Appendix 3: Detailed results for determining the best coefficients

Table 2: Journal coefficient impact on MeSH Ranking (MeSH=1)

Journal Coefficient	0	0.1	0.3	0.4	0.45	0.48	0.49	0.5	0.51	0.52	0.55	0.6	1	2
Median Rank	76	63	57	58	56	57	57	54	57	58	57	57	66	99

Table 3: Study Design coefficient impact on MeSH and Journal Relevance Ranking (MeSH=1, Journal Relevance=0.5)

Study Design Coefficient	0	0.1	0.2	0.25	0.29	0.3	0.31	0.32	0.35	0.4	0.5	0.7
Median Rank	54	51	47	43	43	41	41	42	42	44	46	49

Detecting Associations between Major Depressive Disorder Treatment and Essential Hypertension using Electronic Health Records

Jyotishman Pathak, PhD[1] Gyorgy Simon, PhD[2] Dingcheng Li, PhD[1] Joanna M. Biernacka, PhD[1] Gregory J. Jenkins, MS[1] Christopher G. Chute, MD,DrPH[1] Daniel K. Hall-Flavin, MD[1] Richard M. Weinshilboum, MD[1]

[1]Mayo Clinic, Rochester, MN [2]University of Minnesota, Twin Cities, MN

Abstract

In this observational study, we investigate the correlation between depression and hypertension on a cohort of patients treated for major depressive disorder using Selective Serotonin Reuptake Inhibitors (SSRIs) and assess the effect of depression treatment on the diagnoses and treatment for essential hypertension. Our results indicate that the positive effect of successful depression treatment can be discovered and estimated from electronic health record (EHR) data even for a small sample size. We have also successfully detected differences in the effect of depression treatment in hypertensive patients between the two phenotypes representing successful treatment outcomes—response and remission— concluding that achieving remission has a longer lasting effect than response.

1. Introduction

It is well known that among high utilizers of medical care, a process for systematic identification and treatment of depression is often regarded as the cornerstone to significant and sustained improvements in clinical outcomes, and potentially reducing healthcare related costs. Several efforts[1,2] have explored this within the context of collaborative care management for patients with depression and chronic illnesses, and the early results demonstrate promise in the integration of interventions in real-world practices. However, these studies have been conducted in a controlled clinical trial environment, and the evidence is remains preliminary with regard to the effectiveness of collaborative care in primary care settings and the bi-directional impact on treatment for depression and cardiovascular diseases[3].

To address this critical gap, we conducted an electronic health record (EHR) data-driven observational study on patients who had essential hypertension and major depressive disorder (MDD), and subsequently treated with selective serotonin reuptake inhibitors (SSRIs) to analyze the effect of treated hypertension on treatment response to MDD. In particular, we included the patient cohort (N=794) that was enrolled as part of an 8-week outpatient SSRI clinical trial within the Mayo Clinic Pharmacogenomics Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS; ClinicalTrials.gov number: NCT00613470), and applied structured data as well as natural language processing (NLP) queries to retrospectively extract vital signs, medications, diagnoses, smoking status and other co-morbidities for hypertension from the patient EHRs. We developed a linear mixed effect model to associate the success of depression treatment with improvement in hypertension control, and our results indicate that the positive effect of successful depression treatment can be discovered and estimated from EHR data even for a small patient cohort (N=135 with hypertension out of 794 depressed patients). We have also successfully detected differences in the effect of depression treatment in hypertensive patients between the two phenotypes representing successful treatment outcomes—response and remission—arriving at the conclusion that achieving remission has a longer lasting positive effect on treated hypertension than response. We acknowledge these findings are preliminary and provide an early insight in associating MDD treatment response with essential hypertension, but nonetheless demonstrate the applicability of secondary use of EHR data for answering an important question that has significant implications in improved patient outcomes and reducing the healthcare burden.

2. Background

2.1 Major Depressive Disorder and Hypertension

Depression is a risk factor for hypertension[4,5], and studies have shown an association with poor compliance with anti-hypertensive treatment regimens. However, studies investigating the association between high blood pressure (BP) and psychopathology have not produced consistent results, primarily for two major classes of psychiatric ailments: MDD and anxiety. Some have shown increased BP among patients with depression[6], whereas others have found no association[7], and even in some cases, a decrease in the BP measurements for depressed patients[8]. A possible explanation for this lack of consensus could be that antidepressant use confounds the relationship between psychopathology and BP. For example, antidepressants such as Venlafaxine increase adrenergic activity which leads to higher BP. Similarly, Serotonin (5HT) can cause constriction or dilatation in various vascular systems. In a prospective study of patients treated with antidepressants[9], those who took an SSRI had a 78% increased chance of being prescribed blood pressure medication compared with those who did not. In addition, several clinical trials[1,3] have tried to shed more light on the effect of SSRIs on hypertension. While the findings are still inconclusive and inconsistent—few show an increase in BP and others demonstrate the converse.

DEMOGRAPHICS	PGRN N=794	
	N	%
Gender		
Female	492	62.0%
Male	302	38.0%
Age		
18-30	283	35.6%
31-50	341	42.9%
>51	170	21.4%
Race		
White	739	93.1%
Black or African American	13	1.6%
Other	42	5.3%
Education*		
<High School	30	3.8%
High school but < college degree	239	30.2%
> College degree	523	66.0%
Employment		
Employed	605	76.2%
Retired	44	5.5%
Unemployed	145	18.3%
Marital Status*		
Divorced	127	16.0%
Married	376	47.4%
Never	274	34.6%
Widowed	16	2.0%
***% of totals are figured based data available**		
College degree is considered greater than or = to 14 yrs of schooling		

Table 1 Demographics for the PGRN-AMPS cohort

Further, apart from the TrueBlue[3] study by Morgan and colleagues, to our knowledge, none of them have investigated this within a primary care setting either prospectively or retrospectively leveraging patient data from EHR systems. The focus of our study is the latter, and in particular, investigating the correlation between essential hypertension and treatment response to SSRIs for patients diagnosed with MDD using data from EHRs.

2.2 Mayo Clinic Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS)

The Mayo Clinic Pharmacogenomic Research Network Antidepressant Medication Pharmacogenomic Study (PGRN-AMPS[10]) is an NIH funded study that is investigating the pharmacogenetics of SSRI treatment response to MDD. The study was designed as an 8-week outpatient SSRI clinical trial performed at the Mayo Clinic in Rochester, MN, USA. Patients enrolled in the study met diagnostic criteria for MDD without psychosis or mania and had a 17-item Hamilton Depression Rating Scale (HAMD-17[11]) score >=14. The study was designed with inclusion and exclusion criteria similar to those used in the Sequenced Treatment Alternatives to Relieve Depression study (STAR*D[12]). Specifically, potential study subjects taking an antidepressant, antipsychotic or mood-stabilizing medication were excluded. Patients with MDD initially received either 10 mg of escitalopram or 20 mg of citalopram. SSRI efficacy and treatment response was determined using the 16-item Quick Inventory of Depressive Symptomatology (QIDS-C16[13]) scores after 4-weeks and then 8-weeks of SSRI therapy. The patients were further followed-up by the study team at 24 weeks. At 4-weeks after the initiation of treatment, the dose could be increased to 20 mg of escitalopram or 40 mg of citalopram after a clinical assessment of the subject. All patients provided written informed consent. The study protocol was approved by the Mayo Clinic Institutional Review Board. The two primary outcomes that we investigate in this study are "response" (defined as $\geq$ 50% reduction in QIDS-C16 score from baseline to the last visit) and "remission" (defined as a QIDS-C16 score of $\leq$ 5 at the last visit). For both outcomes, we do separate analysis using linear mixed modelling for building predictive classifiers.

3. Materials and Methods

3.1 Cohort and dataset description

The PGRN-AMPS cohort is comprised of 794 Mayo Clinic patients (see **Table 1**). The study population contained more females (62%) than males (38%), and was overall younger with only 21% above the age of 51 at the time of enrollment. For all study participants, baseline, 4-weeks, and 8-weeks HAMD-17 and QIDS-C16 scores were recorded. At each visit, if the QIDS-C16 score was $\leq$5, indicating remission, the dose would be maintained. If the score was between 6 and 8, a clinical decision was made to either maintain or increase the dose. A QIDS-C16 score of $\geq$9 would lead to a dose increase unless contraindicated. The specific dosing information along with the depression scores were adequately captured in the PGRN-AMPS dataset. In addition to the information collected as part of the trial, we retrospectively applied structured data and NLP queries using cTAKES (http://ctakes.apache.org) for extracting vital signs (including systolic and diastolic BP measurements), medications, diagnoses, signs and symptoms and smoking status from the patients' EHR at Mayo Clinic.

3.2 Predictive modeling and classification

The data studies the longitudinal trends (defined as an increase or decrease over time) in BP measurements in essential hypertension for a number of patients (N=135) treated for depression whose antidepressant treatment outcomes can be grouped into either remission or response. In our EHR data, multiple observations exist for these patients both "before" and "after" the trial. We assume that trends before the trial ("*pre-trial trends*") are common across all patients regardless of depression outcome, but after the beginning of the trial, the trends ("*post-trial trends*") may change. We capture the patient's pre-trial trends in a predictive model that we call the "*pre-trial trend model*". The pre-trial trend model is a linear mixed effect model clustered by patient (random effect) that predicts the BP measurements based on two fixed effects: time (relative to the beginning of the trial) and depression outcome for treatment response. With hypertension being a chronic disease (most likely under control), we consider 3 years as a relatively short period of time. We built our pre-trial model on data observed during the 1.5 year period preceding the beginning of the trial and make predictions using the model for time periods ranging between 6 months and 3 years after the beginning of the trial. Since the mechanism underlying the depression outcome can also affect the BP measurements, we included the depression outcome as a fixed effect. This setup allows for each patient to have his own baseline BP measurement and an effect based on his depression treatment outcome.

Once the pre-trial trend model is developed, we use it to make a prediction for BP measurements within a certain window (of at most 3 years after the beginning of the trial). We compare this predicted measurement with the observed result from patient's EHR data. The residual (difference) is in part noise, and in part, a systematic bias that stems from the pre-trial trend model's inability to correctly model the post-trial trend. Since we attribute the change in the trends (from pre- to post-trial) to the depression treatment, the systematic bias quantifies the effect of the depression treatment on the BP measurements. For patients, who did not see any improvement in their depression status, this bias should be close to 0. But for patients, who saw improvement in their depression status and this improvement in depression was accompanied by an improvement in hypertension, the bias is positive. To assess the significance of the bias, we applied bootstrapping with the sampling unit being a patient (rather than the observation). In other words, we created bootstrap samples of the data set by sampling the patients with replacement and including *all* observations pertaining to the selected patient (possibly multiple times). We used 2000 bootstrap iterations to estimate p-values and confidence intervals.

4. Results and analysis

From the PGRN-AMPS study, we identified 135 patients with depression and essential hypertension. For these patients, we collected relevant covariates including age, gender, race, marital status, smoking status and history of obesity from the EHR data along with all blood pressure measurements. We also had these patients' QIDS-C16 scores at baseline (beginning of the trial), 4 and 8 weeks into the trial at our disposal. As the treatment regimen could be adjusted at 4 weeks, we decided to only use the scores at baseline and at 8 weeks. We used the QIDS-C16 scores to define our depression treatment phenotypes following the PGRN study. The population is divided into three groups: patients whose depression remitted, patients who showed response (improvement) but whose depression did not remit, and patients with effectively no response. Formally, patients with a QIDS-C16 score of less than 5 at week 8 form the "remission" phenotype, patients whose QIDS-C16 score dropped by half between baseline (beginning of the trial) and 8 weeks form the "response" phenotype and the remaining patients who did not experience significant improvement are the "controls". Based on this definition, out of 135 patients, 83 remitted, 17 responded and 35 were control patients. The remission and response phenotypes both indicate (at least partially) successful depression treatment, but we consider them separately.

Our primary clinical interest lies in the effect of depression on essential hypertension. Specifically, we aim to assess whether the SBP for patients who underwent successful treatment is lower than it would be if the treatment had been unsuccessful. To achieve this, we applied the pre-trial trend model to predict the SBP level at 180 days and also at 3 years (1080 days). The importance for these time points is as follows: At 180 days, the non-linear trends that likely exist in the data still have only negligible effect on the predictions from the pre-trial model, and hence, we can relatively accurately quantify the effect of successful depression treatment. However, 180 days is indeed a short time period and it may not be sufficient for the successful treatment to significantly impact blood pressure. At 3 years, we have sufficient time to realize any gains that the successful depression treatment may have conferred on the blood pressure. This effect may be diluted (i.e., depression may have recurred) and the pre-trial trend model may get increasingly inaccurate the further we move away from (with respect to time-periods) the beginning of the trial.

Figure 1 depicts the residuals of the control, response and remission patients for all observations obtained during a window of 180 days (in the left pane) and 1080 days (in the right pane). A time window of 180 days starts 8 weeks after the beginning of the trial and ends 180 days later. Recall that the residual is the (signed) difference between the observed SBP and the expected SBP under the assumption that the pre-trial trend continued after the beginning of the trial. This assumption is tantamount to saying that the treatment had no effect. The residuals are signed, such that they reflect *improvements*: positive residuals indicate that the observed SBP was *lower* than what we expected. There is a residual for each observations, and thus the figure presents the residuals summarized into boxplots, one boxplot for each depression treatment phenotype: control, response and remission. Comparing the median residuals (the tick horizontal lines in the middle of the three boxes) there is a substantial difference between control and response patients reflected in the window size at 180 days after the beginning of the trial, and remission is "catching up" by day 1080. This tendency may continue for beyond 1100 days (i.e., beyond 3 years from the beginning of the trial), however, we cannot ascertain this because the pre-trial model becomes increasingly invalid resulting in excessive estimation errors.

Our results in **Figure 1** suggest not only that successful depression treatment affects blood pressure (and thus hypertension) beneficially, but also that a difference exists between the two phenotypes representing successful depression treatment. To further illustrate this difference, we depict the mean improvement (reduction in SBP) for the two phenotypes

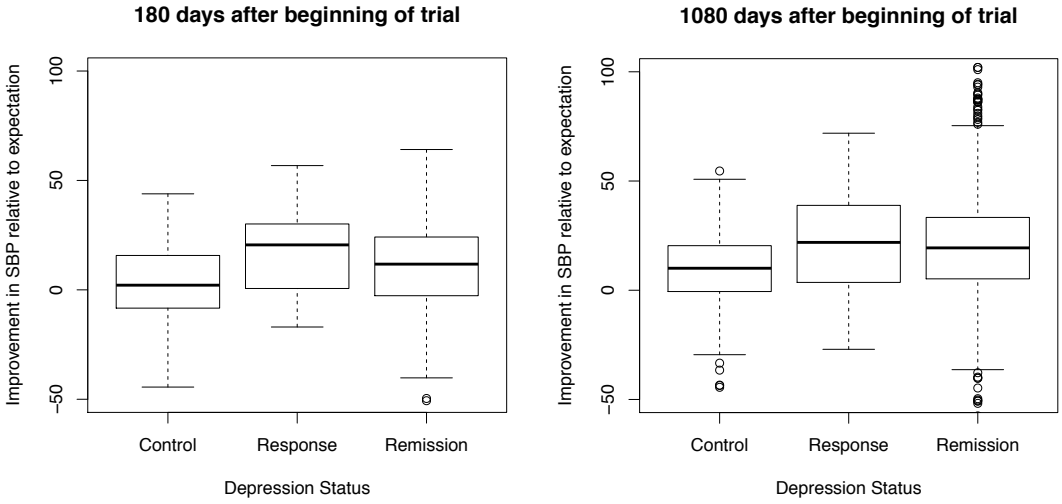


Figure 1: The residuals (reduction in SBP) over 180 days (left pane) and 1080 days (right pane) after the beginning of the trial summarized as boxplots. Significant reduction in SBP has been achieved through successful depression treatment.

over increasingly long time windows between 180 and 1100 days (at 30 day increments). All time windows start 8 weeks after the beginning of the trial. **Figure 2** depicts the mean of the residuals and their confidence interval (vertical axis) as a function of time window size (horizontal axis). For example, in case of the 180-day window, the mean of the residuals is approximately 13, meaning that on average, response patients saw SBP levels of approximately 13 mmHg less than expected. The mean as well as the confidence interval was obtained through bootstrap estimation from 2000 replications.

The left pane presents the results for response patients (**Figure 2**). Recall that depression treatment successfully reduced the QIDS-C16 score by half for response patients; these patients are still clinically depressed (their QIDS-C16 score still exceeds 5). **Figure 2** also shows that this substantial reduction in QIDS-C16 score was accompanied with an almost immediate improvement in SBP with a lasting effect: their SBP is lower than what we would expect without

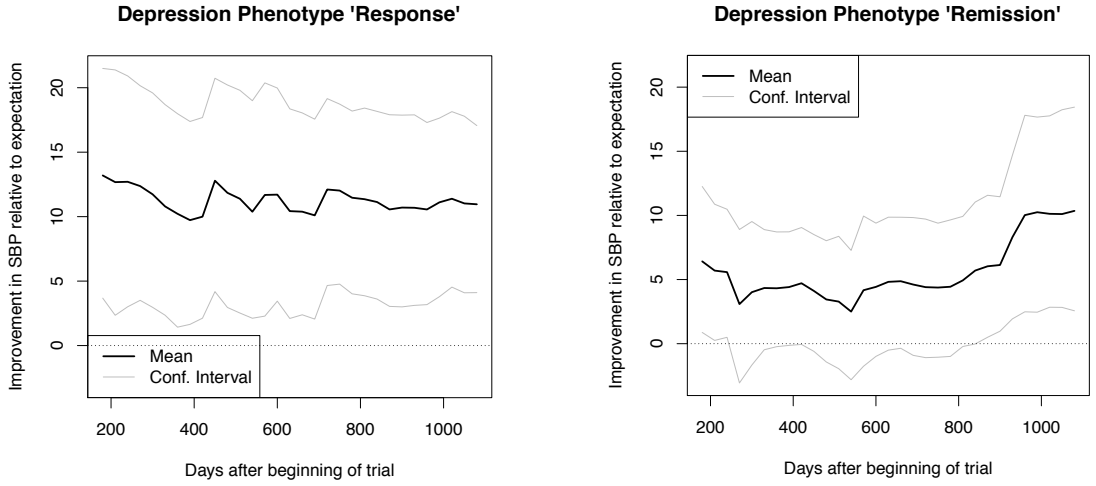


Figure 2: The mean improvement (reduction in SBP) in the Response and Remission phenotypes during windows of 180, 210, …1100 days after the beginning of the trial. This illustrates the difference in SBP response between the Response and Remission phenotypes.

successful depression treatment as long as three years after the trial. The effect is significant throughout the 1100 days. Remission patients had a baseline QIDS-C16 score of at least 7 and by week 8, it reduced to a score of 5 or below. Relative to the response patients, the reduction in QIDS-C16 score can be more modest. However, these patients are no longer clinically depressed. The effect of depression treatment in this phenotype can only be seen later. It is only after 2.5 years where the effect of this phenotype is significant. It is interesting to point out the difference between the response and remission phenotypes: for response, the effect of depression treatment on SBP is immediately noticeable, but decreases

over time. However, for remission, the effect of depression treatment on SBP slowly builds up over time. This could suggest that remission has a more lasting effect on SBP than response.

5. Discussion

5.1 Summary

Acknowledging that factors associated with depression such as sedentary activity or unhealthy nutrition may predispose an individual to hypertension underscores more fully the relationship between depression and hypertension and will assist in the management of both conditions. While several clinical trials in the recent past have attempted to evaluate the bidirectional relationship between depression and hypertension with appropriate real-world interventions, in this study, we illustrate the feasibility of studying such a relationship retrospectively using an ambulatory EHR on out-patients who were diagnosed with MDD, and were subsequently treated with SSRIs. We found that the positive effect of successful depression treatment can be discovered and estimated from EHR data even for a small sample size (N=135 patients diagnosed for MDD and essential hypertension). We have also successfully detected differences in the effect of depression treatment in hypertensive patients between the two phenotypes representing successful treatment (response and remission) arriving at the suggestion that achieving remission has a longer lasting effect than response.

5.2 Limitations

We took great precaution in estimating the significance of the treatment effects. We performed simulations to estimate probabilities and confidence intervals, which was necessary due to the clustered nature of the data set. However, we need to point out, that with only 7 "response" patients for hypertension, the subpopulations we worked with are very small and may not be representative of the general population. The modeling approach we took effectively eliminates (or reduces) individual variability of the patients, but we did not have sufficient sample size to estimated additional fixed effects such as age, gender, and ethnicity. Besides, contributing to the individual variability (which we largely accounted for) these factors may also influence the pre-trial trend. When the proposed approach is applied to an entire EHR, rather than a small subset of the cohort that was used in a clinical trial, these limitations are virtually eliminated. In fact, the sheer number of patients in an EHR may allow for non-linear modeling of the trends—a topic that we plan to explore in future.

5.3 Future work

As mentioned above, one of the major limitations of our work is the smaller cohort size. Hence, our immediate plans are to experiment with the linear mixed model in a larger cohort of patients diagnosed with MDD at Mayo Clinic. Further, we intend to investigate EHR data-driven analysis for understanding the association for treatment response to MDD with multiple chronic illness as well as specialty care, including surgeries. The objective here would be to ascertain the bi-directionality of the relationship with respect to treatment response to these conditions. In collaboration with the Mayo Clinic PGRN, our eventual goal is to understand better the pharmacogenomics of antidepressant treatment response and its impact on differing health care outcomes.

5.4 Conclusion

In this retrospective study, we investigate the correlation between depression and hypertension on a cohort of patients treated for MDD using SSRIs and assess the effect of depression treatment on the diagnoses and treatment for essential hypertension. While limited by the size of our cohort, our preliminary results provide positive evidence on the impact of response to antidepressant therapy for patients with hypertension.

Acknowledgment. This research is supported in part by the Mayo Clinic Early Career Development Award (FP00058504), PheMA (R01-GM105688), SHARP (90TR002), PHONT and PGRN (U19-GM061388) networks.

References

1. Bogner HR, de Vries HF. Integration of Depression and Hypertension Treatment: A Pilot, Randomized Controlled Trial. *The Annals of Family Medicine.* 2008;6(4):295-301.
2. Katon WJ, Lin EHB, Von Korff M, et al. Collaborative Care for Patients with Depression and Chronic Illnesses. *New England Journal of Medicine.* 2010/12/30 2010;363(27):2611-2620.
3. Morgan MAJ, Coates MJ, Dunbar JA, Reddy P, Schlicht K, Fuller J. The TrueBlue model of collaborative care using practice nurses as case managers for depression alongside diabetes or heart disease: a randomised trial. *BMJ Open.* 2013;3(1).
4. Michal M, Wiltink J, Lackner K, et al. Association of hypertension with depression in the community: results from the Gutenberg Health Study. *Journal of Hypertension.* 2013;Publish Ahead of Print.
5. Wu E-L, Chien IC, Lin C-H, Chou Y-J, Chou P. Increased risk of hypertension in patients with major depressive disorder: A population-based study. *Journal of Psychosomatic Research.* 9// 2012;73(3):169-174.

6. Rutledge T, Hogan BE. A Quantitative Review of Prospective Evidence Linking Psychological Factors With Hypertension Development. *Psychosomatic Medicine.* 2002;64(5):758-766.

7. Shinn EH, Poston WSC, Kimball KT, St. Jeor ST, Foreyt JP. Blood pressure and symptoms of depression and anxiety: a prospective study*. *American Journal of Hypertension.* 2001;14(7):660-664.

8. Hildrum B, Mykletun A, Stordal E, Bjelland I, Dahl AA, Holmen J. Association of low blood pressure with anxiety and depression: the Nord-Trøndelag Health Study. *Journal of Epidemiology and Community Health.* 2007;61(1):53-58.

9. Monte S, Macchia A, Romero M, D'Ettorre A, Giuliani R, Tognoni G. Antidepressants and cardiovascular outcomes in patients without known cardiovascular risk. *Eur J Clin Pharmacol.* 2009/11/01 2009;65(11):1131-1138.

10. Ji Y, Biernacka JM, Hebbring S, et al. Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. *Pharmacogenomics J.* 08/21/online 2012.

11. Hamilton M. Development of a rating scale for primary depressive illness. *Brit. J. Soc. Clin. Psychol.* 1967;6(1):278-296.

12. Rush AJ, Fava M, Wisniewski SR, et al. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials.* 2// 2004;25(1):119-142.

13. Doraiswamy PM, Bernstein IH, Rush AJ, et al. Diagnostic utility of the Quick Inventory of Depressive Symptomatology (QIDS-C16 and QIDS-SR16) in the elderly. *Acta Psychiatrica Scandinavica.* 2010;122(3):226-234.

Variation in Cohorts Derived from EHR Data in Four Care Delivery Settings

Susan Rea, PhD[1], Kent R. Bailey, PhD[2],
Jyotishman Pathak, PhD[2], Peter J. Haug, MD[1]
[1]Intermountain Healthcare, Salt Lake City, UT; [2]Mayo Clinic, Rochester, MN

Introduction

EHR data are desirable for secondary use in health research but are known to have inconsistencies. The context of their origination, or provenance, may affect the comparability of EHR data for secondary usage. We investigated suspected differences in demographic and comorbidity data and availability of information among four health care settings: ambulatory office visits, hospital inpatients, emergency visits, and visits for tests and other diagnostic and treatment procedures only. Descriptive comparative results suggest that cases accrued to a diabetes cohort in these settings differed on demographics, morbidity profiles and completeness of EHR data. The distribution of cases among the four settings also differed by provider geographic regions. These differences may reflect real differences in the health care services and documentation practices as well as differential patient access and utilization among the settings. This study demonstrates generalizable methods of classifying encounters in order to profile, compare and inform the use of secondary data across organizations.

Background

Hripcsak and Albers discuss the need for a better understanding of EHR data in context of the primary use environment where data were generated. This knowledge will enable innovative solutions to deal with known biases in secondary data.[1] Richesson, et al., acknowledged that phenotyping algorithms developed in particular health care settings will have unique biases in patient characteristics, utilization of services, and documentation practices that affect the performance of the algorithm in other settings.[2] Jensen, et al., describe the 'overfitting' and systematic bias in EHR data as a significant weakness in the pursuit of machine learning and prediction from secondary health care data. [3]

However, studies exposing specific effects of EHR data provenance were not found. Understanding these effects and their generalizability across organizations enhances the usability of secondary EHR data. This analysis was part of an ongoing collaborative study of the effects of heterogeneity of EHR data on the generalizability of a Type 2 diabetes mellitus (T2DM) phenotyping algorithm.[4] [5] This algorithm was selected as a use case to demonstrate the SHARPn data normalization pipeline.[6] The pipeline persists normalized secondary EHR data for analytic processing, using standard terminologies and detailed clinical element models (CEMs).[7] CEMs support rich provenance, or context, for each data element. This analysis shows differences in the profiles of patient cohorts related to the context of the setting of care. We used standardized administrative data to classify the settings of care so that the analysis can be replicated in other health care delivery organizations.

Methods

Intermountain Healthcare is a nonprofit, integrated health care system with 22 hospitals, 185 ambulatory clinics, and a full spectrum of health services such as home care, rehabilitation, laboratories and advanced trauma centers. Its services cover the state of Utah. Electronic health record (EHR) systems have been used since before 1983.[8] There are two administrative systems: records of patient encounters for professional services by the medical group, clinical laboratories, and other ambulatory services; and records of institutional encounters, including hospitals, hospital-based clinics, and other sites of care such as nursing home and home health. Longitudinal EHR and administrative data for secondary use are managed in the Enterprise Data Warehouse (EDW). These resources enable the study of provenance factors on a large heterogeneous repository of patient data. In this study, we focused on demographic and comorbidity features of a patient cohort with evidence of DM, compared by the type of practice settings they visited. Intermountain IRB approval was granted for this study.

Approximately 10% of 110,000 adult patients with evidence of DM during 2007 – 2011 were randomly sampled from the EDW (n = 10,426). One ICD-9-CM code for DM (250.*) in the encounter diagnosis records in either the institutional or professional services administrative systems was used as the selection criterion for a potential DM case, irrespective of type. Hospital discharge diagnoses and ambulatory encounter diagnoses in all ICD-9-CM

sequence positions were included. All study data were drawn from this same 5 year period. We used standard administrative coding of CMS *place of service* (POS) [9] for professional services, institutional patient types (inpatient, outpatient) and emergent arrival status, and Berenson-Eggers Type of Service (BETOS) [10] codes to classify all encounters into four health care delivery setting groups. The encounter settings used in the study were (1) face to face provider visits where evaluation and a medical diagnosis are expected to occur, (2) hospital stays, (3) hospital emergency room visits, and (4) encounters for tests and procedures only. The differences in characteristics of the patients were compared for those who had at least one provider evaluation visit; those who had no known evaluation visit but had either inpatient or, separately, emergency encounters; and those who had none of the previous types of encounters but had visits for tests and procedures. The four comparison groups are referred to as (1) *AMB*, (2) *IP not AMB*, (3) *ED not AMB*, and (4) *TP Only*.

Nationally standard administrative data were used to classify the setting groups in order to generate generalizable results. Only four POS codes were used to define eligible professional service encounters: 11 (office), 20 (urgent care facility), 22 (outpatient hospital) and 81 (independent laboratory). Institutional administrative systems must categorize encounters as *inpatient* or *outpatient* and whether *emergent*. The BETOS codes summarize Healthcare Common Procedure Coding System (HCPCS) codes into six major groupings: physician evaluation and management (E&M), physician procedures, imaging, laboratory tests, durable medical equipment, and other. All HCPCS, which include CPT4, codes available in either administrative system were mapped to BETOS codes. We used the combination of POS 11, 20, or 22 or institutional outpatient (non-emergent) and a BETOS code of M1A, M1B or M6 [11] (ambulatory E&M or consultation services) to define the 'AMB' group. 'IP not AMB' and 'ED not AMB' groups were based on institutional patient types only. The 'TP Only' group consisted of POS 81 or institutional outpatient (non-emergent) type and a BETOS code of P*, I* or T* (procedures, imaging, tests).

Diagnosis data were summarized in Clinical Classifications Software[12] single-level categories in order to reduce many ICD-9-CM codes into meaningful groupings for analysis. Several CCS categories that are known comorbidities to DM were selected to compare for this analysis: DM complications, hypertension, coronary heart disease and chronic renal failure. Patient deaths noted in the EDW through July, 2013, were used. Death data were updated from the Utah state records in April, 2013. Intermountain has assigned its facilities to physical regions of the state of Utah and southeastern Idaho. For this analysis, the regions were summarized further into the *urban* regions, the *rural* areas, and *mixed* regions – those having local access to health care facilities and resources but distant from the urban centers. Visit counts were summarized both to compare utilization across settings and as a proxy measure of the depth of information that may be expected in the EHR.

Patient demographics, visit counts, and comorbidities were described for the 4 groups to assess whether patients identified as diabetics retrospectively from EHR data appear to differ by the setting of care.

Results and Discussion

Table 1 shows the distribution of the 10,426 cases to the health care delivery setting groups. The majority of patients with a DM diagnosis code in the 5 year period were seen at least once for medical evaluation (77%). The 'AMB' group showed a slightly higher proportion of women as well as much higher average visit counts. The 'IP not AMB' group was older, on average. The number of visits varies among the groups. There were 112 cases (1.1% of the study cohort) that had no visits classified into the groups used for comparison. The unclassified cases contained a higher proportion of men than other groups and even less visit history. The encounters for these cases were reviewed using additional administrative data. They comprised lab or imaging (40%), hospital general clinic visits (31%), hospital specialty clinic visits (17%), and professional services in other settings (13%). About half of these could be classified by incorporating local administrative data into the classification methods, but the intent was to use standardized administrative data to generate the groupings.

The distribution of setting groups in each region type is shown in Figure 2. We measured the percentage of all encounters in each region type that were classified to each setting. Each region's four setting percentages total 100. Regions may share cases so these data can only suggest a trend in access to care in these settings. The trend is for a higher proportion of cases in the 'AMB' setting in urban versus rural regions (75% v. 52%) and a higher proportion of DM cases in the 'ED not AMB' setting (23% v. 11%) and the 'TP Only' setting (18% v. 9%) in the rural versus urban regions.

VISIT SETTING	# CASES	% OF COHORT*	AVG AGE	% FEMALE	AVG VISIT COUNT	AVG DM VISIT COUNT
AMB	7984	76.6	57.6	51.4	31.7	9.6
IP not AMB	972	9.3	61.9	47.3	8.4	3.0
ED not AMB	1192	11.4	56.9	47.7	7.6	2.9
TP only	583	5.6	57.8	46.8	5.2	2.0
unclassified	112	1.1	58.1	41.1	2.4	1.5
* IP and ED rows share 417 cases.						
** Known deaths updated July, 2013.						

Table 1. Distribution of cases and characteristics by setting

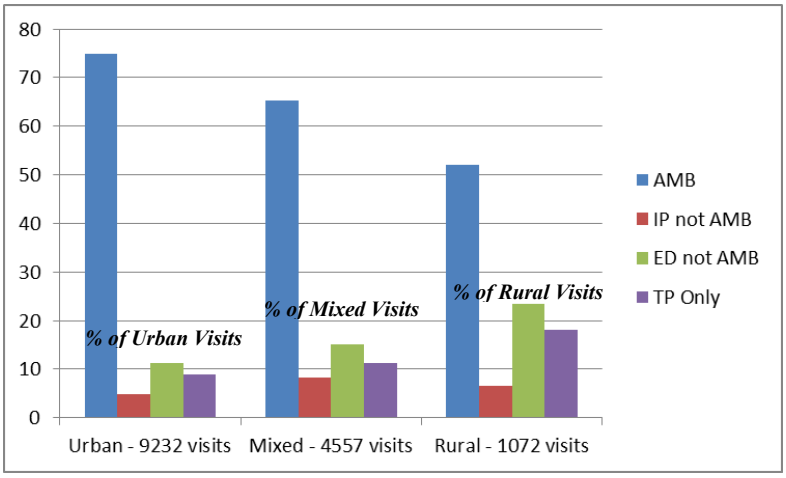


Figure 1. Percent of cases accrued to setting groups by regions

The volume and settings of visits were compared across the setting groups (Table 2). The 'IP not AMB' group had more inpatient encounters than other groups. We cannot compare ED visits between the 'IP not AMB' and the 'ED not AMB' groups because standard hospital coding assigns inpatient status to patients admitted from the ED. Table 2 confirms the visit data in Table 1 showing few encounters over 5 years for all settings other than 'AMB'. A 43% rate of hospitalization in the 'AMB' group was similar to a rate of 42.6% previously reported for 18,404 diabetes cases.[13]

VISIT SETTING	% HAVING IP VISITS	AVG# IP VISITS PER CASE	% HAVING ED VISITS	AVG# ED VISITS PER CASE	% HAVING TP VISITS	AVG# TP VISITS PER CASE
AMB	43	1	50	2	94	14
IP not AMB	100	2	43	1	63	3
ED not AMB	35	1	100	2	55	3
TP only	0	0	0	0	100	4

Table 2. Visit volume across settings by setting groups

Figure 2 shows a potential problem with mixing the cases drawn from these settings in a secondary research cohort. The proportion of documented comorbidities and known death per case are shown for each setting group. The mortality rates are highest in the 'IP not AMB' group and suggest this group would have more morbidity. Although the proportion of cases with coronary heart disease and chronic renal failure are somewhat higher than the 'AMB' group (34% v. 27% and 21% v. 15%), the hypertension and DM complications proportions follow the decreasing trend for comorbidities for all settings compared to the 'AMB' group. These data suggest there may be less documentation for comorbidities related to the coding practices in the non-'AMB' settings or related to the setting groups' lower average visit counts shown in Table 1.

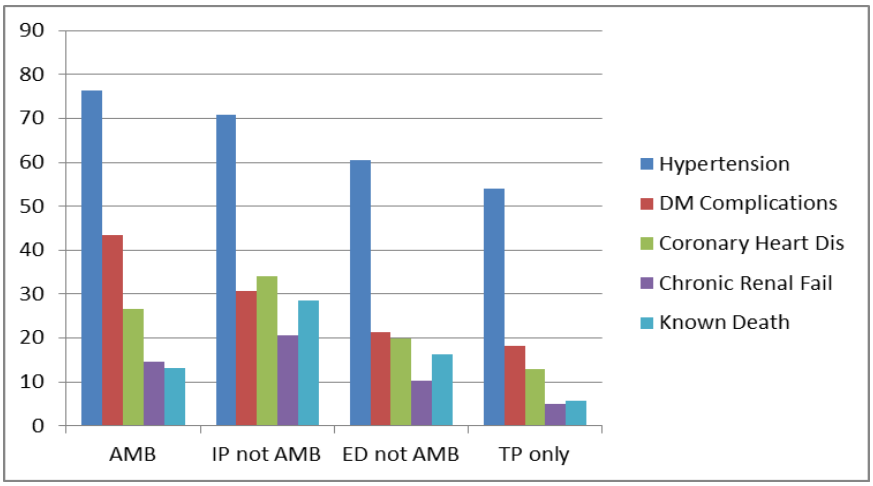


Figure 2. Proportion of cases having documented comorbidities or death

Conclusions

These data suggest demographic, morbidity, mortality and data availability differences in DM cases by provider setting where they might be identified by a DM phenotyping algorithm. The comorbidity profile for cases in setting groups having fewer encounters probably reflects more missing data rather than lower disease burden. Intermountain, like other health care delivery organizations, has a unique mix of provider settings and population access to the settings. This descriptive study was intended to discover and describe setting data differences that might affect the comparability and usability of secondary EHR data. We also demonstrated standard methods to classify cases into the setting groups so that similar profiling may be used to compare cohorts across organizations. Although administrative data, included in an EHR, were used for this study, they signal downstream effects in the clinical observations recorded in these settings. The provenance of clinical observations in the EHR also conveys important primary use contextual information that can inform the comparability of secondary data aggregated for research.

Limitations and Further Research

The classification of institutional outpatient data may be improved by the use of standard revenue codes in addition to or in place of BETOS codes. These data were a 'snapshot' of cases over a 5 year time period, with consequent loss of case data outside these bounds. The data were sampled from one organization and results do not reflect other provider organizations. The focus of our research is not the explanation of specific inconsistencies in EHR data, but rather to contribute generalizable methods to expose data quality issues and remediation opportunities. Further

research of data provenance and data heterogeneity across provider organizations and the effects on the accuracy of specific phenotyping algorithms is underway.

Acknowledgements
This manuscript was made possible by funding from the Strategic Health IT Advanced Research Projects (SHARP) Program (90TR002) administered by the Office of the National Coordinator for Health Information Technology. The contents of the manuscript are solely the responsibility of the authors.

References
1 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inform Assoc. 2013 Jan 1;**20**(1):117-21.
2 Richesson RL, Rusincovitch SA, Wixted D, et al. A comparison of phenotype definitions for diabetes mellitus. J Am Med Inform Assoc. 2013 Sep 11.
3 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature reviews Genetics. 2012 Jun;**13**(6):395-405.
4 Bailey KR, Rea S, Wood-Wentz CM. Extracting Data from EHRs for Algorithm Implementation at Two Comprehensive Care Institutions Data Sources, Pitfalls, Uncertainties. AMIA Joint Summit on Clinical Research Informatics; 2013; San Francisco, CA; 2013.
5 Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. J Am Med Inform Assoc. 2012 Mar-Apr;**19**(2):212-8.
6 Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. Journal of Biomedical Informatics. 2012;**45**(4):763-71.
7 Coyle JF, Mori AR, Huff SM. Standards for detailed clinical models as the basis for medical data exchange and decision support. Int J Med Inform. 2003 Mar;**69**(2-3):157-74.
8 Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. Journal of medical systems. 1983 Apr;**7**(2):87-102.
9 Place of Service Code Set. 2012 Nov, 2012 [cited July, 2013]; Available from: http://www.cms.gov/Medicare/Coding/place-of-service-codes/Place_of_Service_Code_Set.html
10 Berenson-Eggers Type of Service (BETOS). 2012 Dec 6, 2012 [cited June 1, 2013]; Available from: http://www.cms.gov/Medicare/Coding/HCPCSReleaseCodeSets/BETOS.html
11 Berenson-Eggers Type of Service (BETOS) Codes. 2012 [cited Jul, 2013]; Available from: http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/MedicareFeeforSvcPartsAB/downloads/betosdesccodes.pdf
12 Clinical Classifications Software (CCS) for ICD-9-CM. 2013 Mar 28, 2013 [cited June 1, 2013]; Available from: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp
13 Robbins JM, Thatcher GE, Webb DA, Valdmanis VG. Nutritionist Visits, Diabetes Classes, and Hospitalization Rates and Charges: The Urban Diabetes Study. Diabetes care. 2008 April 1, 2008;**31**(4):655-60.

Enhancing Electronic Health Records to Support Clinical Research

David K. Vawdrey, PhD[1], Chunhua Weng, PhD[1], David Herion, MD[2], James J. Cimino, MD[3,1]

[1]Columbia University Department of Biomedical Informatics, New York, NY; [2]Department of Clinical Research Informatics, NIH Clinical Center, Bethesda, MD; [3]Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD

Abstract

The "Learning Health System" has been described as an environment that drives research and innovation as a natural outgrowth of patient care. Electronic health records (EHRs) are necessary to enable the Learning Health System; however, a source of frustration is that current systems fail to adequately support research needs. We propose a model for enhancing EHRs to collect structured and standards-based clinical research data during clinical encounters that promotes efficiency and computational reuse of quality data for both care and research. The model integrates Common Data Elements (CDEs) for clinical research into existing clinical documentation workflows, leveraging executable documentation guidance within the EHR to support coordinated, standardized data collection for both patient care and clinical research.

Introduction

The separation of research from patient care processes has long been a barrier to achieving the goals of the "Learning Health System"[1,2] and makes clinical research unnecessarily time-consuming and expensive. The Institute of Medicine called for increased attention to this problem in Crossing the Quality Chasm,[3] and the National Research Council lamented that IT-related activities of health professionals are "rarely well integrated into clinical practice," and that health IT is "rarely used to link clinical care and research."[4]

Electronic health records (EHRs) have long been viewed as a catalyst to expedite clinical research. Multiple institutions, such as Kaiser, the VA, Partners HealthCare, and Intermountain Healthcare, have been using EHRs to support clinical research for decades.[5] Mayo Clinic conducts more than 4,000 clinical trials each year, and nearly every trial relies on EHR information.[6] The Cleveland Clinic has been using their EHR for trial recruitment,[7] as has Stanford University.[8] The University of Texas MD Anderson Cancer Center developed ClinicStation that presents integrated views of data from both patient care and research.[9] Despite these examples, clinical research is often poorly integrated with clinical care. Poor integration results in unnecessary duplication of work and limits learning from clinical practice.[10]

Besides using an EHR system, many hospitals with large clinical research programs have implemented clinical trial management systems (CTMS), which maintain administrative and clinical information of research participants and are usually disconnected from EHRs. The design requirements for CTMS and EHR systems differ significantly, especially with respect to their data models. While EHRs are oriented to single-patient, unplanned care-related tasks, CTMS tools are designed to support protocol-based research tasks. Another distinction is that EHRs typically contain information obtained through what may be considered "routine data collection" (i.e., data collected in the process of providing clinical care), while CTMSs may require "specialized data collection" (i.e., data collected with finer granularity or more precision). For example, a routine blood pressure measurement recorded in the EHR in a hospital may not be suitable for inclusion in a clinical research trial dataset because the patient's body position (e.g., sitting, standing, supine) was not documented. In this case, a research nurse would be obliged to take a separate blood pressure measurement, recording the value and the body position in the CTMS or directly in a research case report form (CRF). Similar examples of redundancy in clinical and research tasks can be seen with ordering of laboratory and imaging tests.[11]

Several groups have published on the barriers to integrating clinical and research information systems.[12-17] To address some of these issues, interoperability standards have been developed for exchanging data between clinical and research systems. The Clinical Data Interchange Standards Consortium (CDISC) has established global, vendor-neutral, and freely available standards to support the acquisition, exchange, submission and archive of clinical research data and

metadata. Partnering with CDISC, members of the Integrating the Healthcare Enterprise (IHE) initiative have labored to link EHR and clinical research systems through efforts such as the Retrieve Form for Data Capture Profile (RFD), which allows clinical trial forms to be embedded in EHRs and pre-populated with certain data. With RFD, no data are retained in the EHR, which is a significant drawback if the data are useful for clinical as well as research purposes. Overall, adoption of clinical research data exchange standards remains low. According to CDISC, which counts over 200 organizations in its worldwide membership, barriers to adoption include a lack of understanding of the relevant standards, the cost of implementation, and the lack of data for exchange.[18]

Many clinical research institutions have adopted Vanderbilt University's Research Electronic Data Capture (REDCap) software, or similar electronic data capture (EDC) systems to store research data. In some cases, a CTMS system is used for research subject tracking, billing, and visit/procedure scheduling, and a separate EDC system is used to store research data. While these systems have been thoughtfully designed to accommodate a variety of data import and export options, extracting EHR data from proprietary vendor data models and synchronizing information across multiple systems remains challenging. In the end, far too much effort is spent working around the limitations of EHRs as opposed to addressing the underlying challenges.

In this paper, we propose a model for enhancing EHRs to collect structured and standards-based clinical research data during clinical encounters that promotes efficiency and computational reuse of quality data for both care and research. The model integrates Common Data Elements (CDEs) for clinical research into existing clinical documentation workflows, leveraging executable documentation guidance within the EHR to support coordinated, standardized data collection for both patient care and clinical research.

Methods

Process and Limitations of Electronic Documentation

Spurred by government financial incentives, the United States is experiencing unprecedented adoption of EHRs, including increasing use of electronic clinical documentation. Electronic documentation improves legibility and availability of notes, and it facilitates the collection of structured data for purposes such as quality improvement and research. However, implementing electronic documentation has been reported to adversely impact clinicians' perceptions of documentation quality, workflow, professional communication, and patient care.[19-22]

The question of "What should be documented in the EHR?" is relevant and timely. The 2011 AMIA Invitational Health Policy Meeting addressed the current and future state of technology-enabled clinical data capture and documentation, and in February 2013, the Office of the National Coordinator for Health Information Technology (ONC) HIT Policy Committee's Meaningful Use and Certification and Adoption workgroups held hearings focused on clinical documentation functionality in EHRs and its effect on the delivery of high quality clinical care and provider efficiency and collaboration. Even after decades of experience with EHRs, electronic notes continue to be cluttered and redundant, making it difficult for clinicians to understand the actions and thought processes of their colleagues.[19] Our attempt to enhance EHR documentation capabilities to support clinical research acknowledges that the primary purpose of clinician documentation must be to support patient care.

Common Data Elements (CDE)

Frequently, data collection forms used in clinical research contain fields with inadequate definitions and idiosyncratic permissible values.[23] Common Data Elements (CDEs) have been developed with the goal of reducing the time and effort spent by researchers deciding what data to collect for a clinical trial, as well as increasing the interoperability of data collected by various groups. An example of a CDE is shown in Figure 1. CDEs are defined in detail using a

metadata dictionary and can be shared in a standardized format across multiple institutions. Our model for enhancing the EHR to support clinical research integrates CDEs with current electronic documentation workflows.

The use of CDEs is a growing trend, although to date, adoption has occurred on a relatively small scale—most commonly in cancer research.[24-27] The National Cancer Institute (NCI)'s Cancer Data Standards Registry and Repository (caDSR) supports development and deployment of CDEs in cancer research and provides a web-based CDE Browser and application programming interface for public use.[28] CDEs have also been used in epilepsy research,[29] posttraumatic stress disorder research,[30] traumatic brain injury,[31] and substance use disorder.[32] Recently, the National Institute of Neurological Disorders and Stroke (NINDS) has "strongly encouraged" investigators of Phase III and exploratory clinical trials to use CDEs.[33] Additionally, the National Library of Medicine has become an important participant in curating CDEs for clinical research and recently developed the CDE Resource Portal (http://www.nlm.nih.gov/cde/). In the 2013 AMIA Joint Summit meeting, Lin et al. described a method for mapping the clinically-oriented Common Element Model to research variables in dbGaP.[34] The work presented a useful taxonomy of contextual information to be recorded when collecting research data. Our model goes beyond definition to represent clinical workflows, and to create an environment to collect research data during clinical encounters.

Data Element Details

Public ID:	2435448
Version:	1.0
Long Name:	Cigarette Consumption Daily Count
Short Name:	CIG_COMPN_D_CT
Preferred Question Text:	During periods when you smoked, how many cigarettes did you or do you usually smoke per day?
Definition:	The number of cigarettes consumed per day.
Value Domain:	Cigarette Consumption Daily Count
Data Element Concept:	Smoking Use
Context:	PS&CC
Workflow Status:	RELEASED
Origin:	
Registration Status:	Qualified
Direct Link:	https://cdebrowser.nci.nih.gov/CDEBrowser/search?elementDetails=9&FirstTimer=0&PageId=ElementDetailsGroup&publicId=2435448&version=1.0

Figure 1. Example of a Common Data Element (CDE) definition for capturing smoking history.

Integrating CDEs with Documentation Workflows

Figure 2 presents a model that integrates CDEs with existing documentation workflows to improve the process of collecting data for clinical research. The model, which emphasizes clinician and researcher data needs and documentation processes, is informed by the conceptual framework for clinical research informatics proposed by Kahn and Weng.[35] Our model consists of an informatics-enabled clinical research workflow, where providers or clinicians can access a library of disease-specific CDEs and perform CDE-based structured data collection using smart templates. Documentation decision support can guide clinicians in capturing research-quality data. In this implementation, the EHR plays a dual role for both patient care and clinical research and facilitates the interoperability of the processes of both missions.

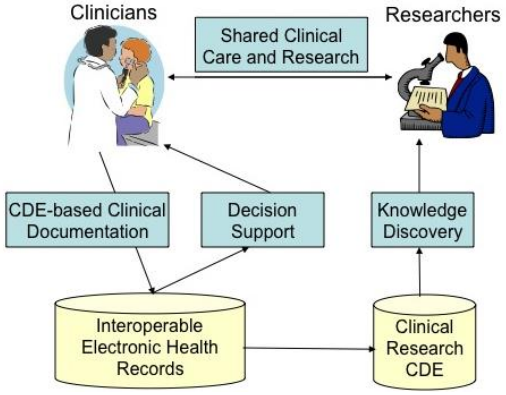


Figure 2. A conceptual model that integrates CDEs with existing EHR documentation workflows.

Implementation

The following scenario highlights the EHR's potential to enhance clinical research, as well as the challenges that may be overcome by our proposed model for enhancing EHR documentation to support clinical research.

Ms. Johnson is a 52-year-old woman who arrives in the emergency department complaining of abdominal pain, fatigue, and jaundice. Reviewing the results of her blood work, physicians discover that Ms. Johnson has tested positive for hepatitis C. She is admitted to the hospital. During the intake process, Ms. Johnson's nurse asks about

her health history, including her smoking status. Ms. Johnson quit smoking 3 years ago when her granddaughter was born, but prior to that time she had smoked approximately half-a-pack of cigarettes per day since she was a teenager.

Two years before Ms. Johnson's hospital visit, the institution had implemented a certified EHR and attested to its "meaningful use" of the system, qualifying to receive federal incentive payments under the HITECH act. As part of its EHR/meaningful use implementation, the hospital configured a structured "Nursing Admission History" documentation template. The template contained check-boxes to record smoking status, with options such as "Never," "Current," and "Former."

Consulting with her physicians and family members, Ms. Johnson elects to enroll in a phase 3 clinical trial of a new Hepatitis-C medication called sofoviran. The clinical trial's purpose is to confirm the effectiveness and safety of the drug, and the clinical trial sponsor agency is interested in collecting detailed information of several types, including the health history of trial participants as well as any adverse events they experience—such as headaches or chest pain—that could be caused by the medication.

One of the data elements that must be recorded in the pharmaceutical company's case report form (CRF) is the study participant's smoking history. The CRF specifically requires documentation about the level of cigarette consumption (i.e., packs-per-day) for current and former smokers. Because this level of granularity is not captured in the EHR, a research nurse must re-ask the patient about her smoking habits.

Several hours after Ms. Johnson receives her first dose of sofoviran, she develops a severe headache. The headache is a possible adverse event of the medication, and should be recorded on the CRF. She describes the headache to her physician during evening rounds, and the doctor informally notes the pain in his free-text assessment/plan without identifying the possible connection to the medication.

Aspects of this scenario are probably familiar to clinicians and research investigators in a variety of environments. Applying the model in Figure 2 can expose the overlap between data elements collected during routine patient care and data elements that are captured as part of a specific research protocol. Current documentation workflows can then be augmented by mapping EHR data fields to CDEs. For example, referring to the above scenario, smoking status recorded for every hospital patient can be encoded in a computationally reusable format such as the CDE for "Cigarette Consumption Daily Count" shown in Figure 1.

Moreover, the proposed model will fuse documentation workflows with awareness of clinical research protocol documentation requirements such as recording of adverse events. Applying the model to the scenario above, a decision support algorithm could prompt the physician during the note-writing process to report possible adverse events using standard definitions. If an adverse event is identified, it can be coded using a CDE and appropriately communicated to the trial sponsor and other stakeholders. For serious adverse events, CDE concepts can be leveraged to generate the necessary codes, forms, and messages (e.g., MEDWATCH, ICH E2B, ICSR) for transmission to systems such as the FDA's Adverse Event Reporting System (FAERS). Similarly, the decision support system can provide the EHR with temporal context that is crucial for most types of clinical research (e.g., alerting the clinician that a certain panel of laboratory tests must be performed during the third week of a protocol, or allowing the clinician to tag the test results as being the "week 3" results).

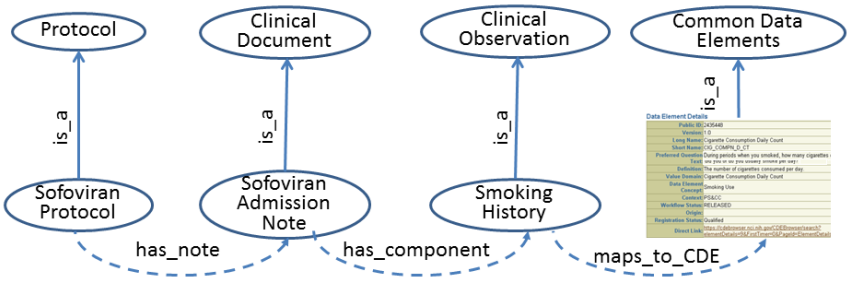


Figure 3. Representation of relationships between clinical research and EHR documentation concepts for a hypothetical "Sofoviran" clinical trial protocol.

Significant effort will be required to fully implement the proposed model for improving EHR documentation processes; however, institutions with certified EHR systems are much closer to achieving the vision of a learning heath system than they were just a few years ago. Figure 3 illustrates how the data collection model can be encoded in a terminology management system as a set of concepts, including CDEs (with their various attributes), EHR observations that correspond to particular CDEs (such as "smoking history CDE"), EHR clinical documents (such as "Sofoviran Admission Note"), and clinical trial

protocols (such as "Sofoviran Protocol"). We intend to implement this model both in the Research Entities Dictionary[36] at the NIH Clinical Center and the Medical Entities Dictionary at Columbia University Medical Center.[37]

High-level classes can be defined to represent concepts that need to be mapped between the clinical and research realms—not only data definitions, but also the data workflows and context necessary for computational reuse. For EHR concepts, classes include the data source and the EHR data context. Intermediate-level concepts within those classes include 1) for data source: clinical narrative note, clinical structured template, clinical flow sheet, laboratory test, registration data; and 2) for EHR data context: time of data collection, visit the data are attached to, status (final, preliminary), linked clinical order. For clinical research concepts, classes include the data type hierarchy and the research data context. Intermediate-level concepts may include symptoms, signs, laboratory tests, diagnoses, and procedures. The research data context includes concepts that define the constraints on research data collection that are usually executed by the research staff that carry out a trial. Examples include the time that data are collected (either absolute time with respect to entering a clinical trial or relative to other trial events), allowable sources (e.g., only values measured in a special laboratory, or only diagnoses confirmed by a physician), and other constraints (e.g., measurements taken after a meal).

In our model, documentation decision support can be encoded as an open source set of computer-interpretable process rules for coordinating clinical care and research workflows to facilitate knowledge sharing. The result of a rule firing can be the automatic addition of data elements to a template that is about to be used, a message to a user, or some other type of decision support. Given the similarity between clinical guidelines and process rules guidelines, it makes sense to leverage existing standards for clinical guidelines to formalize the process rules. Many languages have been developed to represent and share formal knowledge of research protocols or clinical guidelines, such as the Arden Syntax,[38] GELLO,[39] PROforma,[40] EON,[41] GLIF,[42] and SAGE.[43]

Discussion

There are compelling arguments for integration of patient care and clinical research, both in terms of workflow processes and electronic systems.[44,45] A recent decision support panel identified four areas where advances in decision support lie: the state of the knowledge base (the set of rules, content, and workflow opportunities for intervention); necessary database elements to support decision support functions; operational features to promote usability and to measure performance; and organizational structures to help manage and govern current and new decision support interventions.[46] The panel's findings stress the central importance for decision support functions and workflow changes to be mutually supportive to each other so that decision support facilitates workflow changes and relies on workflow support and integration. Mandl and Kohane emphasized the value of flexibility in healthcare system design, arguing that "system[s] will have to function under evolving policies and in the service of new health care delivery mechanisms…and emerging information technologies."[47] Their SMART platform enables lightweight, modular "apps" to be integrated with EHRs, overcoming the proprietary "silos" that exist in current systems.[48] As this system architecture paradigm gains momentum, EHR implementers will be increasingly in a position where the 'right choice' in terms of designing data collection forms is also the 'easy choice'—flexible and efficient user interfaces will enable clinicians to capture discrete, coded data that are computationally reusable.

Our proposed model for enhancing EHRs to support clinical research builds on the foundation of CDE standards, bridging the adoption gap by incorporating them directly into electronic documentation tools in the EHR. The model facilitates reuse of routinely collected data and seamless inclusion of data capture specific to a patient's research studies while minimizing the impact on clinician effort. The model is consistent with next-generation EHR architectures such as the SMART platform, enabling documentation decision support within the EHR to support coordinated, standardized data collection for both patient care and clinical research.

Conclusion

The clinical research informatics community has emphasized the need for innovative information technology to support clinical and clinical research processes; however, the complexity of the patient care and clinical research environments makes coordination among the multiple stakeholders difficult to achieve. We propose a model for enhancing EHRs to collect structured and standards-based clinical research data during clinical encounters that promotes efficiency and computational reuse of quality data for both care and research. While we believe that the model will be useful in a variety of healthcare delivery settings, further research is warranted to demonstrate its effectiveness.

References

1. Weng C, Appelbaum P, Hripcsak G, et al. Using EHRs to integrate research with patient care: promises and challenges. *J Am Med Inform Assoc.* Sep-Oct 2012;19(5):684-687.
2. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med.* Nov 10 2010;2(57):57cm29.
3. Institute of Medicine (U.S.). Committee on Quality of Health Care in America. *Crossing the quality chasm : a new health system for the 21st century*. Washington, D.C.: National Academy Press; 2001.
4. Stead WW, Lin H, National Research Council (U.S.). Committee on Engaging the Computer Science Research Community in Health Care Informatics. *Computational technology for effective health care : immediate steps and strategic directions*. Washington, D.C.: National Academies Press; 2009.
5. FastCures. Think Research: Using EMR to Bridge Patient Care and Research. 2009; http://www.fastercures.org/objects/pdfs/white_papers/emr_whitepaper_summary.pdf.
6. Cimino JJ, McNamara TJ, Meredith T, et al. Evaluation of a proposed method for representing drug terminology. *Proc AMIA Symp.* 1999:47-51.
7. Cimino JJ, Patel VL, Kushniruk AW. Studying the human-computer-terminology interface. *J Am Med Inform Assoc.* Mar-Apr 2001;8(2):163-173.
8. STRIDE; http://clinicalinformatics.stanford.edu/STRIDE/. Accessed April 14, 2013.
9. ClinicStation. http://www.informatics-review.com/wiki/index.php/ClinicStation. Accessed March 16, 2013.
10. *Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions* National Research Council;2009.
11. El Fadly A, Daniel C, Bousquet C, Dart T, Lastic PY, Degoulet P. Electronic Healthcare Record and clinical research in cardiovascular radiology. HL7 CDA and CDISC ODM interoperability. *AMIA Annu Symp Proc.* 2007:216-220.
12. Kuchinke W, Wiegelmann S, Verplancke P, Ohmann C. Extended cooperation in clinical studies through exchange of CDISC metadata between different study software solutions. *Methods Inf Med.* 2006;45(4):441-446.
13. Kuchinke W, Aerts J, Semler SC, Ohmann C. CDISC standard-based electronic archiving of clinical trials. *Methods Inf Med.* 2009;48(5):408-413.
14. Bruland P, Breil B, Fritz F, Dugas M. Interoperability in clinical research: from metadata registries to semantically annotated CDISC ODM. *Stud Health Technol Inform.* 2012;180:564-568.
15. Dworkin RH, Allen R, Kopko S, et al. A standard database format for clinical trials of pain treatments: an ACTTION-CDISC initiative. *Pain.* Jan 2013;154(1):11-14.
16. El Fadly A, Lucas N, Rance B, Verplancke P, Lastic PY, Daniel C. The REUSE project: EHR as single datasource for biomedical research. *Stud Health Technol Inform.* 2010;160(Pt 2):1324-1328.
17. El Fadly A, Rance B, Lucas N, et al. Integrating clinical research with the Healthcare Enterprise: from the RE-USE project to the EHR4CR platform. *J Biomed Inform.* Dec 2011;44 Suppl 1:S94-102.
18. (http://www.cdisc.org/stuff/contentmgr/files/0/fdf8540f5324c81f48d3630923b95fd6/misc/cdisc_journal_friggle_etal_p2.pdf) Accessed March 13, 2013.
19. Embi PJ, Yackel TR, Logan JR, Bowen JL, Cooney TG, Gorman PN. Impacts of computerized physician documentation in a teaching hospital: perceptions of faculty and resident physicians. *J Am Med Inform Assoc.* Jul-Aug 2004;11(4):300-309.
20. Hartzband P, Groopman J. Off the record--avoiding the pitfalls of going electronic. *N Engl J Med.* Apr 17 2008;358(16):1656-1658.
21. Hirschtick RE. A piece of my mind. Copy-and-paste. *JAMA.* May 24 2006;295(20):2335-2336.
22. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes: An evaluation of input errors. *Methods Inf Med.* 2003;42(1):61-67.
23. Grinnon ST, Miller K, Marler JR, et al. National Institute of Neurological Disorders and Stroke Common Data Element Project - approach and methods. *Clin Trials.* Jun 2012;9(3):322-329.
24. Jiang G, Solbrig HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. *J Am Med Inform Assoc.* Jun 2012;19(1e):e129-136.
25. Jiang G, Solbrig HR, Chute CG. Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network. *J Biomed Inform.* Dec 2011;44 Suppl 1:S78-85.
26. Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. *Methods Inf Med.* 2006;45(6):594-601.

27. Patel AA, Kajdacsy-Balla A, Berman JJ, et al. The development of common data elements for a multi-institute prostate cancer tissue bank: the Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. *BMC Cancer.* 2005;5:108.

28. Winget MD, Baron JA, Spitz MR, et al. Development of common data elements: the experience of and recommendations from the early detection research network. *Int J Med Inform.* Apr 2003;70(1):41-48.

29. Loring DW, Lowenstein DH, Barbaro NM, et al. Common data elements in epilepsy research: development and implementation of the NINDS epilepsy CDE project. *Epilepsia.* Jun 2011;52(6):1186-1191.

30. Kaloupek DG, Chard KM, Freed MC, et al. Common data elements for posttraumatic stress disorder research. *Arch Phys Med Rehabil.* Nov 2010;91(11):1684-1691.

31. Maas AI, Harrison-Felix CL, Menon D, et al. Common data elements for traumatic brain injury: recommendations from the interagency working group on demographics and clinical assessment. *Arch Phys Med Rehabil.* Nov 2010;91(11):1641-1649.

32. Ghitza UE, Gore-Langton RE, Lindblad R, Shide D, Subramaniam G, Tai B. Common data elements for substance use disorders in electronic health records: the NIDA Clinical Trials Network experience. *Addiction.* Jan 2013;108(1):3-8.

33. http://c-path.org/Events/data-standards-in-clinical-trials/Creating-Common-Data-Elements-for-Neurologic-Diseases.pdf). Accessed March 19, 2013.

34. Lin K, Hsieh A, Farzaneh S, Doan S, Kim H. Standardizing phenotype variables in the database of genotypes and phenotypes (dbGaP) based on information models. Paper presented at: AMIA 2013 Summit on Translational Bioinformatics2013; San Francisco, CA.

35. Kahn MG, Weng C. Clinical research informatics: a conceptual perspective. *J Am Med Inform Assoc.* Jun 2012;19(1e):e36-42.

36. Cimino JJ, Ayres EJ. The clinical research data repository of the US National Institutes of Health. *Stud Health Technol Inform.* 2010;160(Pt 2):1299-1303.

37. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc.* Jan-Feb 1994;1(1):35-50.

38. Hripcsak G, Ludemann P, Pryor T, Wigertz O, Clayton P. Rationale for the Arden Syntax. *Comput Biomed Res.* 1994;27(4):291-324.

39. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. *J Am Med Inform Assoc.* Mar-Apr 2011;18(2):112-117.

40. Fox J, Johns N, Lyons C, Rahmanzadeh A, Thomson R, Wilson P. PROforma: a general technology for clinical decision support systems. *Comput Methods Programs Biomed.* 1997;54(1-2):59-67.

41. Tu S, Musen M. Modeling data and knowledge in the EON guideline architecture. Paper presented at: MedInfo2001.

42. Boxwala A. GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. *Journal of Biomedical Informatics.* 2004;37(3):147-161.

43. Tu SW, Campbell JR, Glasgow J, et al. The SAGE Guideline Model: Achievements and Overview. *JAMIA.* September 1, 2007 2007;14(5):589-598.

44. D'Aunno T. Managing the Care of Health and the Cure of Disease: Arguments for the Importance of Integration. *SO - Health Care Management Review Winter 2001;26(1):85-87.*

45. Marsolo K. Informatics and operations—let's get integrated. *Journal of the American Medical Informatics Association.* September 1, 2012 2012.

46. Teich JM, Osheroff JA, Pifer EA, Sittig DF, Jenders RA, the CDSERP. Clinical Decision Support in Electronic Prescribing: Recommendations and an Action Plan: Report of the Joint Clinical Decision Support Workgroup. *J Am Med Inform Assoc.* July 1, 2005 2005;12(4):365-376.

47. Mandl KD, Kohane IS. No Small Change for the Health Information Economy. *N Engl J Med.* March 26, 2009 2009;360(13):1278-1281.

48. Mandl KD, Mandel JC, Murphy SN, et al. The SMART Platform: early experience enabling substitutable applications for electronic health records. *Journal of the American Medical Informatics Association.* July 1, 2012 2012;19(4):597-603.

Using Software to Elicit User Needs for Clinical Research Visit Scheduling

Chunhua Weng, PhD[1], Mary Regina Boland, MA[1], Yat So, BS[1], Alexander Rusanov, MD[2], Carlos Lopez, MD[3], Richard Steinman, AB[3], Linda Busacca, BA[4], Suzanne Bakken, PhD, RN[1,5], J Thomas Bigger, MD[3]

[1.]Department of Biomedical Informatics; [2.]Department of Anesthesiology; [3.] Department of Medicine; [4.]Clinical Trials Office, [5.]School of Nursing, Columbia University, New York City

Abstract

User needs understanding is critical for developing useful and usable clinical research decision support. Existing methods largely depend on self-reporting and often fail to elicit implicit or fine-grained user needs. We hypothesized that functional software would address this problem by presenting to users existing technology while simultaneously encouraging users to optimize workflow. Using clinical research visit scheduling as an example, we used a piece of software under development that was called IMPACT to reveal user needs iteratively. The identified user needs explained why most clinical research coordinators still rely on paper to schedule clinical research visits. The common user needs themes such as information completeness for software to be useful may generalize to other clinical decision support. This paper contributes valuable firsthand knowledge about user needs for decision support for clinical research visit scheduling among clinical research coordinators and a generalizable methodology for collecting and analyzing software usage data to inform user needs elicitation.

Introduction

Analysis of user needs is necessary in order to develop useful and usable software. Methods, such as focus groups, structured interviews, and ethnographic studies, have been developed for this purpose. However, most of these methods rely on the accounts of intended users based on their experiences with their current work environment so that any new resulting software is more likely to mimic the current work processes rather than offer users a way to explore potential options. A priori requirements engineering also fails to satisfy fine-grained user needs to inform user interface design, whose usability can influence user adoption and user-perceived usefulness of software.

Specifically, user needs understanding is an important problem for the field of clinical research informatics. As Califf pointed out, clinical research sites are the underappreciated components of the nation's clinical research enterprise[1]. Randomized controlled trials (RCTs) are the gold standard for generating high-quality medical evidence. Although Clinical Research Coordinators (CRCs) play a central role in RCTs by coordinating various clinical research activities, they often receive limited technological support[2,3]. This is because their needs for technological support for improving their work efficiency remain largely unknown and therefore unaddressed. Most existing Clinical Trial Management Systems (CTMSs), e.g., Velos eResearch[4] and AllScripts' Study Manager[5], were developed to streamline billing or data management, and thus offer limited support for the workflow of CRCs. The frequent requirement to manage multiple RCTs simultaneously adds to the complexity of the CRCs' workflow.

Of all research activities, visit scheduling is one of the most frequent and time-consuming. To schedule research visits, CRCs must synthesize information from multiple sources, including study calendars, rooms, equipment, and personnel. Despite the availability of scheduling support provided by the existing CTMSs, anecdotally we learned that most CRCs still either rely on paper-based calendaring systems for visit scheduling or perform much manual work around inadequate scheduling software.

We hypothesize that a piece of interactive software could engage users and help them specify their implicit needs thoroughly and hence increase the usability and usefulness of software designed for these users. Through a test of this hypothesis, this paper intends to make two major scientific contributions. First, this paper summarizes the user needs for clinical research visit scheduling decision support and answer this research question, "what is lacking in existing software for clinical research visit scheduling?" Second, the paper illustrates a mixed-methods approach to collecting and analyzing software usage data to help designers understand vague and implicit user needs. This methodology may generalize beyond clinical research visit scheduling to other application domains.

Next we report our experience of using scheduling decision support software to enable CRCs to articulate their implicit and vague user needs for clinical research visit scheduling. Columbia University Medical Center's Institutional Review Board approved this study.

Methods

1. Software Description

To streamline the workflow for scheduling clinical research visits with research resource allocation and optimization, we developed a web-based scheduling system that we called the **I**ntegrated **M**odel for **Pa**tient Care and **C**linical **T**rials (IMPACT)[6]. IMPACT aims to ease the cognitive burden for CRCs for scheduling research visits by automatically synthesizing and computing resource availabilities and recommending suitable dates and times for these visits[6]. CRCs can schedule a research visit, edit task lists, or receive an email or in-system reminder (e.g., "*no breakfast before the screening visit*"). They can also choose from a knowledge base of common tasks when scheduling a research visit. When a CRC schedules a visit, IMPACT presents a protocol-determined target window; IMPACT calculates the duration of the visit from its protocol-specified tasks. IMPACT's resource optimizer presents to the CRC a list of recommended dates and times from which to choose. CRCs can also add PRN (*pro re nata,* Latin for "as needed") visits not specified by protocols[6].

2. Research Processes

Using a previously developed evaluation framework that mixed software log-analysis, a think-aloud protocol, and a survey[7], we recruited CRCs periodically to assess if the software under development meet their user needs. In this paper, we use one recent formative evaluation to illustrate such evaluation processes. In the latest evaluation, we recruited 12 CRCs, 5 men and 7 women with diverse research backgrounds in our institution, to participate in a 30-min scenario-based study section each. Nine CRCs were experienced (2-8 years), while three had between 15 and 20 years of experience. Six CRCs were cardiology specialists, three were behavioral cardiologists two were cancer specialists, and one was a diabetes specialist. The CRCs received no compensation. We asked each CRC to use IMPACT to perform six tasks identified from prior studies of CRCs' workflow[6]: log in, find a patient, schedule a visit, view visit details, reschedule the visit, and update visit statuses. Throughout their 30-min IMPACT session, CRCs were encouraged to talk about their difficulties and "wish lists" regarding IMPACT's interface design and functionalities. We used ATracker[8], an iPad v.2.0 application, to record task-completion-times. Because ATracker records tasks in one-minute units, we recorded all tasks shorter than 1 minute as 0.5 minutes in our analyses. We also counted the steps for performing each task. Furthermore, IMPACT's software log recorded user transitions among the IMPACT screens during each session. We analyzed user action transition frequencies and visualized the results using Cytoscape[9], with directional arrow width indicating the frequencies of transitions.

Results

1. User Action Frequencies

Table 1 shows the frequencies of the user actions.

Table 1. Frequency of use of each function by CRCs, sorted by overall frequency of use

Function	CRC ID												Overall
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
View calendar	34	36	27	30	49	18	25	12	27	15	27	23	323
View visit	9	12	5	9	19	6	8	5	7	5	6	8	99
Calculate time range	4	4	4	6	7	4	5	3	4	3	3	6	53
Log in	1	1	1	4	5	1	1	1	1	2	1	1	20
Schedule visit	3	3	3	4	5	8	5	3	4	3	3	6	50
Reschedule visit	2	1	1	2	3	2	2	1	1	1	1		17
Log out	1	1	1	2	3	2	1	1	1	1	1	1	16
Schedule personal event				1	1								2
View reminder(s)					1		1						2
Change password	1												1
Total	55	58	42	58	93	41	48	26	45	30	42	45	583

Viewing the current calendar was the most frequently used function, followed first by viewing details of individual visits and then by calculating available time ranges for each visit. This result shows that CRCs spent a significant amount of time retrieving information from the calendar; therefore, effective information presentation on the calendar view represents a critical intervention opportunity for improving the efficiency of CRCs during visits scheduling. In contrast, reminders about upcoming visits or their required preparations were rarely viewed. One possible explanation is that CRCs would prefer to receive and read these reminders later when they have more time, rather than when they are not busy using the system.

2. *User Action Transition Graph*

To better understand how CRCs used the most frequently used function, "viewing calendar", we created an action transition graph to show the action transition frequency between action pairs (**Figure 1**). Each node is an action; directional arrow thickness represents the frequency of the action transition. The support for each transition[10] was calculated as the transition frequency divided by the total number of the users (N=12). Numerical labels are shown for arrows with support of at least 1. Users started by logging in and then transitioning from the entry page to the calendar page. A transition with a support of 1 indicates that, on average, each user in the study performed the transition. In general, viewing the calendar was the nexus linking all other actions for visits scheduling. Users navigated from the calendar page to other pages and then back again. The next most frequent action was viewing a visit. After viewing a visit, users typically returned to the calendar. Users also frequently moved between viewing the calendar and interacting with the research resource optimizer. The "resource optimizer" is a feature unique to IMPACT. In contrast, the aforementioned related scheduling software such as Microsoft Outlook, Velos, AllScripts, and WebCAMP support only calendar views and visit views. This feature ranks available time slots for scheduling a research visit[6] based on resource availability and the protocol-determined visit window.

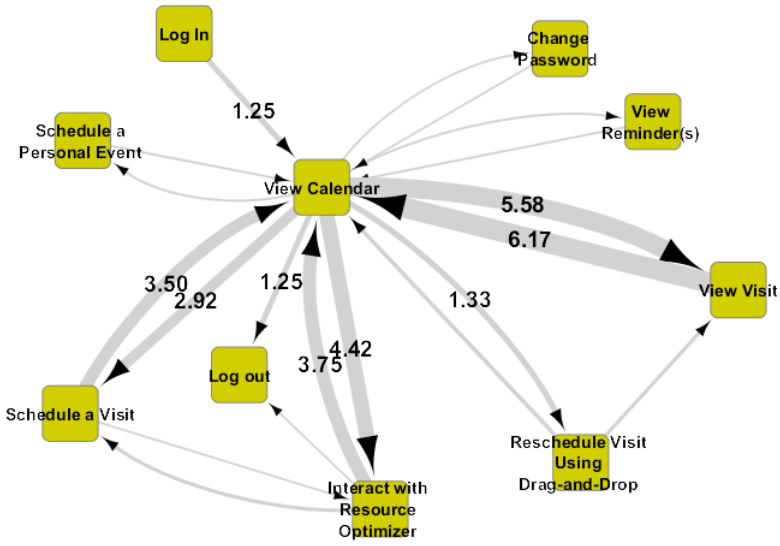


Figure 1. The Action Transition Graph of the usage of the different functions in IMPACT. Each arrow represents a common transition; its thickness indicates the frequency of this transition.

3. *User Need for Standardization across Clinical Trials*

During the "think aloud" session, we were able to capture some user rationales behind certain user preferences for technological support for visits scheduling. We learned that most of the CRCs maintain paper-based systems to schedule visits for their research patients. One such CRC explained that this choice was primarily motivated by the lack of uniformity among sponsor-provided software for visit scheduling and clinical trial management. As is typical, this CRC manages patients for multiple trials with different sponsors simultaneously.

This CRC stated, "***I use pen and paper to schedule my patients…only because the different trials had different methods***. *Different database sets a lot of the times. You're not using the same database to keep their schedule appointments or their entries. So, depending on what day utilize, you sort of attempt to streamline your work according to the particular trial you're working on. So, since it's an--if it's an outside vendor, you just--you are at the emergency, you see what they utilize what they use and then streamline yourself accordingly. So, that's a lot of things.* ***The thing that most vendors and they don't have a uniform set of doing things****. If I'm working with Duke, for instance, they have a way of doing their things. If I'm working with Medtronics, they have a method of their own. And IVRS has their own method****. So, it's never uniform****. Everyone thinks they're doing better than the other guy."*

Another CRC said, *"Yeah. for example, for the month of January, we saw 20 patients. How many patients we--was eligible for being enrolled, which one screen fail or just how many randomization that day for that specific study. If done, that will be great. You know,* ***I don't know how they can lay it out****."*

A third CRC commented, "*What happens when you add another study then? If this--the IMPACT system would have study like ABC and D and then when--then when you're with different tasks,* ***you can have more than one thing going on?...*** *That would be the only thing I would think that, you know, you--and that's the problem.* ***Unless your only job--if your only job is one study, you're not going to have your job for long,*** *you know what I mean? So, you have to be on multiple studies and you have lot of scheduled things. So, I have the studies that have different--…"* [we intercepted], "*maybe it's easier if they had all of these--all this trial information was in there and then you*

could just search the patient and they would already know what trial they were on. We'll probably make scheduling easier if the person called." [The CRC responded,] *"Right, right. So say, one's study, whatever that one I just did. That had all those tasks. Then I have another study that's just--schedule the--a glucose tolerance test and then schedule a radiology procedure that's like four hours in the afternoon that has all these other things that I have to do, you know. So, it would be good to know, you know, what--if it helps you put in that person's study number, then you can have the list and the other list and if things are incomplete."*

Most existing CTMSs are study-specific; however, most CRCs are usually responsible for managing patient scheduling for more than one clinical trial. This explains why many CRCs find it easier to integrate information across patients or across different clinical trials by using a paper-based system. On paper, they can easily draw tables to have an overview of the schedules for all the rooms, personnel, and other research resources. In contrast, aggregating such information electronically is not easy given that different trials have disparate data collection and scheduling software. This finding reveals an implicit user need for supporting cross-study information integration involves multiple trials, CRCs, and patients.

4. User Need for Convenient Information Access and Highlighting

We were able to fine-tune user interface design details based on user feedback, which was only possible by using a software to prompt the users for input. For example, our resource optimizer automatically calculates the time required by a visit by adding up the time required to perform each of the visit's tasks, such as physical exam, blood draw, and mental test. Our software displays time duration in minutes, so that we displayed a duration of 260 minutes for a randomization visit requiring a 30-minute physical exam, a 15-minute blood draw, and a 215-minute mental test. In this case, our testers expressed a preference for the more intuitive display of 4 hours and 20 minutes.

In addition, the users requested other small modifications to the user interface, such as more highlighting: "*The system itself is currently straightforward but just little things like highlighting and stuff would be really helpful. …I didn't realize before that, that when I selected the date, that in the calendar itself it was highlighted, if I had the … would have seen, but normally where I click, I expect that to be highlighted*." Also sometimes users prefer to use mouse to using keyboard, "*How would you do that with the mouse?*"

Furthermore, the testers asked for rationale for the availability of all time slots. Clearly users do not like "black box" reasoning done for them behind the scene but would rather prefer transparency in decision support. We learned these design principles from the users. Before the study, we mistakenly assumed that users prefer to read less information and did not expect them to need explanations for the availability of all time slots.

One CRC described how he would prefer to have everything (e.g., protocol-specific information, patient-specific details, calendar information, etc.) in one screen by stating that, "*One screen with the ability to do multiple entries or dropdown menu allows me to click, and that one screen--let's say there's a dropdown menu here and now I can add this, this, this. It's right on that same menu, you know…If I already have this date here, the screen dates already there. I should be able to dropdown now, add the randomization day."*

"…So, I mean, Schedule, randomization, everything can be done in one shot; don't have to be bouncing back and forth. When this gets full like that, I mean, I could see where it becomes problematic if you got multiple patients and multiple coordinators on one day."

"…(Without IMPACT), the names of the coordinators click, click, click, and all of a sudden, the dates that they're available. (With IMPACT), I can auto write from there the randomization day right from there, the follow-up visit. ***And I don't have to go to multiple places.*** *And then when I say okay, schedule or accept or whatever, you know, bounce me back to the calendar and I could see all the different entries just on the calendar, just to make sure that everything is okay…I could have done this randomization screening is--the screening, there's randomization, there's follow-up visit. Anything that I would have needed to do, I could have done from that very first screening rather than go back to these multiple steps,* ***which is just added time****."*

5. User Need for Fault-tolerant Designs

One common software usability measure is the number of steps (e.g., mouse clicks or page transitions) it takes to complete a task. A lot of applications have purposefully used "one click" feature to speed up the task completion. Our users shared with us their concerns regarding the advantages and disadvantages of enabling "one-click" actions: "*Is there anything that I can lose a lot of information? So I guess not. I hesitate for fear of making mistakes I cannot correct. …If I have to make a lot of changes, obviously, then it's worse; there's a lot of clicking to do. But I try to be very careful when I'm anyway, but if you were to just drop something on it, it wouldn't just delete a ton of events."*

Another CRC had a similar comment when being asked, *"…if you were to improve anything about the IMPACT system, what would it be?*" The answer was "*May be the schedule thing… be able to go back (undo)*."

6. User Need for Separation of Concerns

We became aware of two user needs that we had not previously known, which is that CRCs prefer to have separate time calculation for themselves and for patients and separated calendars for clinical trial tasks and personal events. One CRC stated, "*…our site has a lot of chemotherapy, so those infusions last a really long time.* ***That's going to add a lot of extra time into a visit****. It's great to be able to tell patients that, but in terms of like hours scheduling, it doesn't--because we don't have to do anything toward the infusion so-- I don't know****. Tasks are one thing. This is our time.*** *It's not really****--this is not the patient's time****…. I think that's what I'm having kind of like a weird thing like raping my brain around and this is, like, really more our time, not really the patient's time."*

Another CRC revealed that he created a calendar for every clinical trial study and moved all events related to that study into that calendar so that he could still see all the events, private or work related, but mentally he separated private events from work. IMPACT was designed to read and write into a CRC's calendar to schedule research visits. We confirmed with CRCs that this design was not what they preferred; instead, they prefer IMPACT to read their calendar to know their availability, write to a study-specific shared calendar, and allow them to view this calendar. We would not have been able to detect this user need without the use of the IMPACT as a probing tool to engage users to think about the tradeoffs between information integration and privacy preservation.

7. User Need for Mobility Support

IMPACT was designed as a web-based application. One CRC suggested a mobile version. *"I would say the biggest one that will be easiest for me would be to be able to use the system on my phone as well. Just because also for the portability and be able to look at it and change it as needed, to not only--be bound to this computer can be sometimes cumbersome."*

8. User Needs for Workflow Support

Our previous design for IMPACT allowed users to schedule a single visit but prohibited them from adding subsequent visits until that visit was completed. One user requested the flexibility of scheduling multiple prospective visits simultaneously: "*Is it possible to, like, pre-populate visit, let's say, you know, I have someone who has to come in every three months, if I put their baseline visit in, will it give me tentative visits in the future? And then I can change those to when the patient can specifically come in. That would be really helpful.*"

IMPACT was first designed to generate reminders only for visits that had been scheduled. One asked us to implement a reminder for prospective visits whenever a suitable visit window becomes available, "*So there's not, like, a reminder of when a window comes up?*" We therefore learned the user need for receiving more real-time reminders about potential opportunities to schedule new visits.

Moreover, when there is a cancellation or delay of a scheduled visit, the CRCs also expressed their needs to receive reminders immediately so that they can adjust their appointments accordingly. This is sort of "real-time" plan adaptation that we did not included in our initial design. It turned out that this feature would have very practical utility for CRCs since cancellations or delays are common in clinical research settings and usually headache causes and cost drivers for CRCs.

9. User Need for Information Completeness and Currency

Our test users also provide insights regarding hidden conditions required for a design feature to be useful. For example, one CRC stated, *"Yeah, I do like this integrated calendar feature. This is a--I could get used to that. And it would be nice if everybody like God get their Outlook calendars* ***up to date****, cause then you can see everything."* If users do not update their calendars and make all constraints electronically available, IMPACT will be helpless in terms of synthesizing temporal constraints from user calendars.

Discussion

This paper illustrates how important user needs for clinical research visit scheduling that another method would likely have missed could be acquired by using functional software to prompt users. These user-needs require thorough considerations of a socio-technical approach to engage users and to design useful user interfaces. More importantly, most of these user-requested features are unavailable in existing scheduling software and CTMSs. It was only through insightful input from users that we understood their need for certain views with information

specific to only one user and one trial, and other views with information about many users and trials. Their insights into the incompleteness of information in personal calendars help us to define the best practices expected from users to make IMPACT successful. Their insights into the tradeoffs in "one-click" design also helped us think more about the need for an "undo" option for immediate error rectification. Support for existing behaviors such as using their familiar methods (e.g., a computer mouse) or mobile devices, and scheduling multiple visits, are important. Features that address these needs are being incorporated into the next version of IMPACT, which has evolved from a single-study system, typical of existing CTMS scheduling modules, to a multi-study, multi-CRC collaborative system.

"Paperless office" has been one of the "Holy Grails" chased by computer scientists since early 1980s[11]. However, after several decades, paper proves to be a flexible, extraordinary, and nearly indispensible tool for performing many office tasks[12], including clinical research visit scheduling for clinical research staff. Paper is easy for adding new tables or illustrations, highlighting important information, striking out outdated information, rectifying mistaken information, or moving around and sharing information with different people. Reflecting on the user needs for mobility support and effective information display for IMPACT design, we realized the sophistication of paper use by clinical research coordinators warrants further studies so that we can incorporate implicit user needs or dependency on paper into the design of IMPACT. Any workflow support system design would face adoption obstacles if only mimicking the paper-based workflow or falling short of the existing paper-based system. To win users designers must provide a solution that is superior to the existing paper-based system; otherwise the effort spent for change management would not be worthwhile and the users would not easily buy in the new system.

A common dilemma for clinical decision support or expert systems is "how much information should be presented to users and what should be presented?" Initially we designed the time slot recommendation feature without providing explanations about why certain slots were unavailable under the assumption that users needed solutions more than explanations. The results of this study showed that our assumption was false. The participants in this study taught us that clinical research staff appreciate time-saving advices from expert systems such as IMPACT but also prefer transparency in the logic behind such advices. Therefore, black box decision support may lose users or cost users extra effort to find out "why". Meanwhile, as designers of IMPACT, we are also concerned if the users truly have the time to review all the rationale behind the automatically calculated available slots based on the availability of protocols, personnel, rooms, equipment, patient preferences, and so on many temporal constraints. Therefore, an unsolved question remains, "should user needs be completely defined by users? designers? Or both?"

Since users needs are so complex, one-time user needs elicitation is often insufficient. Reusable methodologies are needed to elicit users needs iteratively. This paper contributes some data collection and analytical techniques for this purpose. The usage log-based action transition graph effectively told us on what tasks users spend more time than others and detect inefficient tasks.

This study has inherent limitations. First, we included a small group of clinical research users from only one institution. A separate study is warranted to test the generalizability of these reported users needs in heterogeneous clinical research settings in different institutions. We believe the knowledge reported here is sufficient to inform the design of user surveys to collect more user needs for research visit scheduling at a larger scale including more institutions. Second, we reported implicit or previously unknown user needs for research visit scheduling but did not demonstrate the clinical impact of such user needs. Ideally data are preferred to show a system design informed by these user-needs leads to better clinical outcomes than a system not informed by these user needs. However, comparative effectiveness research on clinical decision support systems is challenging. Since IMPACT is still going through iterative designs and evaluations, we hope we can validate these user needs when we perform a field trial of IMPACT using real clinical trials and clinical staff and report the results afterwards.

Conclusions

This paper contributes valuable firsthand knowledge about user needs for decision support for clinical research visit scheduling among clinical research coordinators and a generalizable methodology for collecting and analyzing software usage data to inform user needs elicitation. Functional software is a powerful tool and effectively supplements existing methods for eliciting user needs and for arriving at a useful socio-technical design. Future studies can test if these user needs for scheduling clinical research visits may generalize beyond our institution.

Acknowledgments

We thank the CRCs who participated in this study. This study described was supported by grants R01 LM010815 from the National Library of Medicine, R01 HS019853 from the Agency for Healthcare Research and Quality, and UL1 TR000040 from the National Center for Advancing Translational Sciences.

References

1. Califf RM. CLinical research sites—the underappreciated component of the clinical research system. *JAMA.* 2009;302(18):2025-2027.
2. Rico-Villademoros F, Hernando T, Sanz J-L, Lopez-Alonso A, Salamanca O, Camps C, Rosell R. The role of the clinical research coordinator - data manager - in oncology clinical trials. *BMC Medical Research Methodology.* 2004;4(1):6.
3. Khan SA, Kukafka R, Payne PR, Bigger JT, Johnson SB. A day in the life of a clinical research coordinator: observations from community practice settings. *Stud Health Technol Inform.* 2007;129(Pt 1):247-251.
4. Velos eResearch. http://velos.com/solutions/by-product/velos-eresearch-2/. 2013;Accessed January - April 2013.
5. Allscripts Study Manager. 2013;http://investor.allscripts.com/phoenix.zhtml?c=112727&p=irol-newsArticle&ID=816949&highlight=:Accessed January - June 2013.
6. Weng C, Li Y, Berhe S, Boland MR, Gao J, Hruby G, Steinman R, Lopez-Jimenez C, Busacca L, Hripcsak G, Bakken S, Bigger J. An Integrated Model for Patient Care and Clinical Trials (IMPACT) to Support Clinical Research Visit Scheduling Workflow for Future Learning Health Systems. *J Biomed Inform.* 2013 Aug 2013;46(4):642-652.
7. Boland MR, Rusanov A, So Y, Lopez-Jimenez C, Busacca L, Steinman RC, Bakken S, Bigger JT, Weng C. From expert-derived user needs to user-perceived ease of use and usefulness: A two-phase mixed-methods evaluation framework. *Journal of Biomedical Informatics.* 2013(Dec 12. pii: S1532-0464(13)00195-0.).
8. WonderApps-AB. ATracker - Daily Task and Time Tracking Lite. *https://itunes.apple.com/us/app/atracker-lite-daily-task-tracking/id522008611?mt=8.* 2012;Accessed in November 2012.
9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research.* 2003;13(11):2498-2504.
10. Mamykina L, Vawdrey DK, Stetson PD, Zheng K, Hripcsak G. Clinical documentation: composition or synthesis? *J Am Med Inform Assoc.* 2012;19(6):1025-1031.
11. Liu Z, Stork DG. Is paperless really more? *Commun. ACM.* 2000;43(11):94-97.
12. Seong J, Lee W, Lim Y-k. Why we cannot work without paper even in a computerized work environment. *CHI '09 Extended Abstracts on Human Factors in Computing Systems*. Boston, MA, USA: ACM; 2009:4105-4110.

Temporal Analysis of the Usage Log of a Research Networking System

Sunmoo Yoon, RN, PhD[1], Sylvia Trembowelski, MS[2], Richard C Steinman, BA[2], Suzanne Bakken, RN, PhD[1,2,3], Chunhua Weng, PhD[2,3]
[1]School of Nursing, [2]The Irving Institute for Clinical and Translational Research, [3]Department of Biomedical Informatics, Columbia University, New York, New York, USA

Abstract

Despite the proliferation of research networking systems (RNS), their value and usage remains unknown. This study aims to characterize the temporal usage of an RNS, Columbia University Scientific Profiles (CUSP), and to inform the designs of general RNSs. We installed a free usage logging service, Google Analytics, on CUSP and applied time series analysis to compare the usage patterns of the two modes of CUSP: restricted (authenticated) and open access. More users searched by person names than by topics, although the latter enables in-depth vertical search of co-author or co-investigator networks for grants or publications. The open-access mode received more page views but less average time spent on each page than the restricted access mode. The numbers of unique users and searches have increased over the time. This study contributes a trend analysis framework for understanding the usage of RNSs and early knowledge of the usage of an open-access RNS.

Introduction

Facilitating interdisciplinary team science is a critical mission of NIH's Clinical and Translational Science Award (CTSA) program, comprising 60 medical research institutions in 30 states and the District of Columbia in the United States. An important mission of CTSA is to help biomedical researchers identify collaborators. Research Networking Systems (RNS) have been designed in many CTSA institutions to foster collaboration among clinicians and researchers working in multiple disciplines. Understanding information needs of biomedical researchers and collaborator searching methods on RNSs is vital for improving RNSs. One cost-effective way to understand behaviors of biomedical researchers is to analyze web server log files[1]. Log file analysis provides information about system usage, including when, how, where, and by whom the system was used. Although a single method cannot provide a whole picture of user behaviors, previous studies have shown that web usage mining can capture a reasonable amount of information about the performance of a system[2,3].

Columbia University Scientific Profiles (CUSP) (http://irvinginstitute.columbia.edu/cusp) is a locally developed RNS that generates a scientific profile for each biomedical researcher affiliated with the Columbia University Medical Center using information from human resources, MEDLINE databases, and university grants databases. We first launched CUSP in March 2011 for internal use by Columbia University employees. In March 2012, we made CUSP open access. During both phases, we used Google Analytics to monitor its usage in real time.

We previously reported the usage of CUSP by authorized users during its restricted access phase[2]. This study aims to characterize and explain the temporal usages of the open-access CUSP through trend analysis and to gain insights for improving CUSP and other open-access RNSs. Specifically, this study addresses three questions: (1) How has CUSP been used? (2) What are the differences in CUSP usage patterns between restricted access and open access modes? and (3) How has usage changed over time during CUSP's open access period? The Columbia University Medical Center Institutional Review Board approved this study.

Methods

In order to address the first question, we obtained descriptive statistics about CUSP use for the time period from December 2, 2011 to September 19, 2013 from Google Analytics (http://www.google.com/analytics) installed on the CUSP server[4]. The information used for our analysis includes anonymous visitors' geographical locations, Internet service providers, devices, search terms, the number of page views per visit, visitor status (i.e., new or returning), and the number of visits each week, month, and year, as well as overall bounce rates (i.e., percentage of visitors leaving the web site from the home page without performing a search or clicking on any page links). Google Analytics uses tracking cookies to identify unique visitors. Unique searches were recorded along with subsequent profile lookups for scientists, grants, or departments, or co-author or co-investigator network visualization for publications or grants. To detect popular topics searched by CUSP users, we applied content mining to all search terms using Automap (http://www.casos.cs.cmu.edu/projects/automap).

Using Google Analytics timestamp data, we applied time series analysis[5] to address research questions two and three, i.e., to examine the differences of usage trends between different access modes (restricted vs. open versions) and in different years (2012 and 2013). We created time series models using Weka 3.7.9, an open-source machine learning system, to compare the temporal trends between the two access modes and the two time periods. On this basis, we applied the multi-trend regression algorithm based on support vector machine using Weka's SMOreg function[6] to build a trend model for each of the following time periods: restricted access, open access, open-access in year 2012, and open-access in year 2013. We compared the trend model of restricted access with that of open access. We evaluated root-mean-square error (RMSE) to quantify differences between the two access modes and the two time period models, where higher RMSE value indicated more differences. The unit of analysis for research question two was a 100-day period of each access mode: December 2, 2011 through March 11, 2012 for restricted access and March 20, 2012 through June 28, 2012 for open access. For research question three, comparing usage in different time periods during the open access phase, the unit of analysis was a 6-month period to avoid the influence of seasonal changes: March 20, 2012 through September 19, 2012 for open access 2012 and March 20, 2013 through September 19, 2013 for open access 2013.

Results

General Usage of CUSP

During the 21 months from December 2, 2011 through September 19, 2013, 4,974 unique users from 88 countries used CUSP, with a total of 8,492 visits, 28,196 page views, an average of 3.3 pages or 3-minute stay per visit, and a bounce rate of 56%. Half of the visitors (51%) landed directly (4,329 visits) by clicking the CUSP link in the signature section of emails that they received (e.g., faculty signatures that include a link to CUSP). The others (44%) landed through referrals such as the web sites for Columbia University Medical Center, Columbia University College of Physician and Surgeons (3,757 visits), or Google search (378 visits).

Excluding the bounced visitors who left the web site immediately, approximately 60% of the remaining users (2,058 visits) spent between one and three minutes on CUSP, with 20% of them (678 visits) staying for more than 10 minutes. In terms of access methods, 36% of visits came through Intranet within the Columbia University Medical Center, and approximately 7% of users accessed CUSP using a mobile device, such as a smart phone, iPad, or tablet. In terms of the temporal trends of usage, the number of unique visitors spiked to 214 per day between March 20, 2012 and March 22, 2012, when the open-access version of CUSP was released. Afterwards, the number of unique visitors per day stabilized at 10 to 30 during weekdays.

CUSP allows searches for scientists using person names or topics that appear in their publications or grants. A total of 10,210 searches were performed by 49% of all the unique visitors. Most of top 10 searches used person names directly (60%), while the rest (40%) used health topics to retrieve the collaboration network related to the topic. Content mining identified the following 20 most popular topics in **Table 1**.

Table 1. Frequencies of The Top Search topics

Diabetes	212
Obesity	137
Cancer	112
Irving Institute	75
HIV	71
Prevention	64
Heart	62
Informatics	59
Biomedical	47
Cell	46
Health disparities	43
Comparative-effectiveness	42
Ear	40
community	35
Data	34
Cardiology	34
Genetics	29
Breast cancer	29
Pediatric	28
Nursing	28

Figure 1 is an action transition graph generated by ORA (http://www.casos.cs.cmu.edu/projects/ora) using the page access statistics. It illustrates the patterns of the action transitions among the six frequently visited web pages on CUSP. The colors, blue or red, indicate two types of activities grouped by structural similarity in the network as detected by a subgrouping algorithm (CONCOR), which seeks to identify structurally equivalent nodes. The edge width indicates the frequency of transitions between each pair of pages. Users starting from person profile (blue on the top) pages usually reach the publication and grant pages related to a person, whereas users searching by topic (red in the center) reach topic-related grant, publication, or network pages that list names of associated investigators.

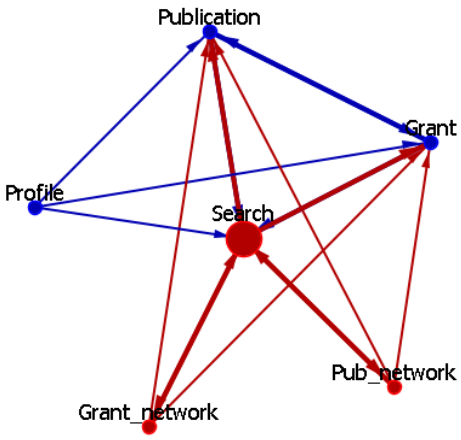


Figure 1. Action transition graph in CUSP (links and nodes sized by path frequency)

Change of CUSP Usage Patterns from Restricted Access (Columbia identifier only) to Open Access

Table 2 compares the usage statistics for the two access modes of CUSP during different time periods. Four times as many unique users visited in the open access mode than in the restricted mode (846 vs. 170). Each user spent 72% less time (11.7 vs. 3.3 minutes) and viewed fewer pages/visit in the open access mode than in the restricted one. A 44% higher bounce rate (38% vs. 55%) was observed in the open access mode. In terms of access, the number of both Intranet and Internet users increased (435% and 52%, resp.). Mobile device usage increased 523% (from 201 to 1,076) after open access.

While both restricted and open-access show a stable number of unique users over time, open access shows a one time spike at the starting point (**Table 2**). Time series analysis shows that the trends of bounces, search refinement and search depth of the two access modes are similar (RMSE<2). Comparing the two modes, trends of visits, unique visitors, and unique searches are only moderately different, with RMSE ranging between 28 and 55. In contrast, patterns of user time spent differ markedly between two modes, with RMSE > 1,000. Using a multi-regression modeling analysis of time series, we were able to plot a forecast model for the number of daily CUSP users of each access mode (**Figure 2**). According to the time series models of each access mode in **Figure 2,** the number of daily users of the restricted access version of CUSP continuously decreased. In contrast, the number of daily users of the open access version has steadily increased.

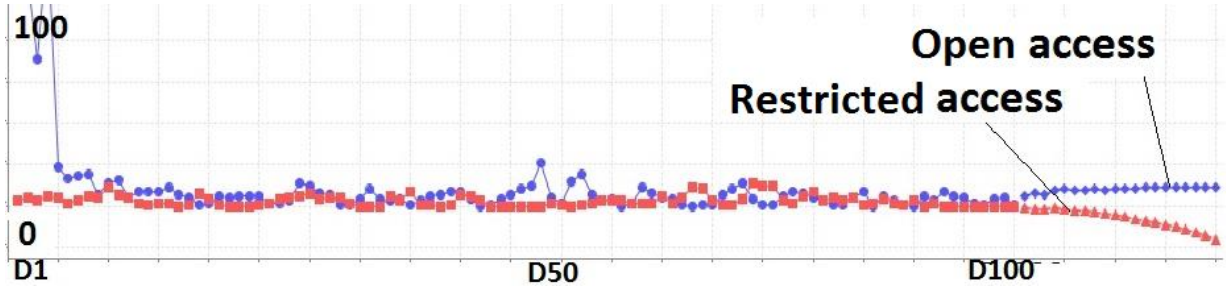
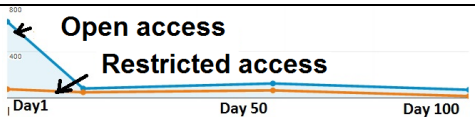
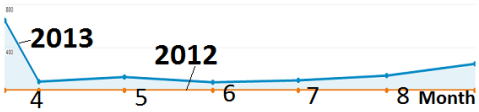


Figure 2. Forecast of the number of daily CUSP users of different access modes using time series analysis

Table 2. CUSP usage in different access modes over time

	Access Mode (unit: 100 days)				Year (unit: 6 months)			
	Restricted access	Open access	Δ (%)	RMSE	2012 Open	2013 Open	Δ (%)	RMSE
Users								
Visits	634	1,735	174	30	2,607	2,244	-14	395
New	170	820	382		1,220	1,631	34	
Returning	464	915	97		1,387	613	-56	
Unique visitors	170	846	398	28	1,245	1,706	37	178
# of Countries	3	14	366		35	64	83	
Engagement								
Duration(min:sec)	11:40	3:20	-72	2,434	3:14	2:29	-23	6,551
New	4:29	2:47	-38		2:49	1:19	-53	
Returning	14:18	3:49	-73		3:36	5:37	56	
Bounce	38%	55%	44	1	56%	57%	2	6
New	58%	46%	-21		47%	63%	34	
Returning	31%	63%	104		63%	42%	-33	
Page/visit	7	4	-51	30	3.4	2.9	-15	37
New	4.2	3.8	-10		3.8	2.4	-37	
Returning	8.2	3.2	-61		3.1	4.4	42	
Access (visits)								
via Intranet	201	1,076	435		1,482	578	-61	
Via Internet	433	659	52		1,125	1,666	48	
Desktop	621	1,654	166		2,499	2,035	-19	
Mobile or tablet	13	81	523		108	209	94	
Incoming traffic								
Direct	220	1,728	214		2,229	613	-72	
Referral	0	6	100		339	1,482	337	
Search	0	1	100		39	149	282	
Content								
Page views	4,514	6,043	34	203	8,973	6,546	-27	3,482
Search								
Unique search	1,002	2,689	168	55	3,656	2,326	-36	754
% refinement	32%	46%	42	1	48%	50%	4	7
Search depth	1.5	0.7	-51	2	0.17	0.16	-6	6
Trends Unique users								

Changes of CUSP Usage Patterns in 2012 and 2013 in Open Access

Table 2 shows that although the number of returning visits decreased substantially in 2013, the numbers of new and unique users steadily increased. The users spent less time viewing slightly fewer pages per visit in 2013 than in 2012. Bounce rate remained similar in 2013 despite the substantial improvement in bounce rate among returning users. In terms of access, while Intranet users decreased, Internet users increased coming from more countries. The mobile or tablet use increased 90% in 2013. The trend graph shows that the number of unique users was higher in 2013 than in 2012. Time series analysis shows that bounce, search refinement, and search depth trends demonstrate similar patterns in 2012 and 2013 (RMSE $\leq$7). However, the trends related to user time spent, page views and unique search revealed completely different patterns between 2012 and 2013 (RMSE>700).

Figure 3 illustrates the overall usage flow of CUSP in 2012 and 2013. The numbers on the flow chart arrow in Figure 3 compare the number of unique searches in 2012 and 2013. While the number of CUSP users coming through institutional website and general search increased, the users from direct approaches (e.g., link in email) decreased in 2013 compared to 2012. Also **Figure 3** shows that there were fewer unique searches for every page in

2013 than there were in 2012. Furthermore, this figure shows that the co-author or co-investigator networks of grants or publications were less utilized than individual profile pages. The numbers on the arrow pointing publication and grant network pages on the right corner show that less than 2% of unique searches (40 out of 2,389) reached the network visualization page, suggesting that CUSP's visualization feature is underused.

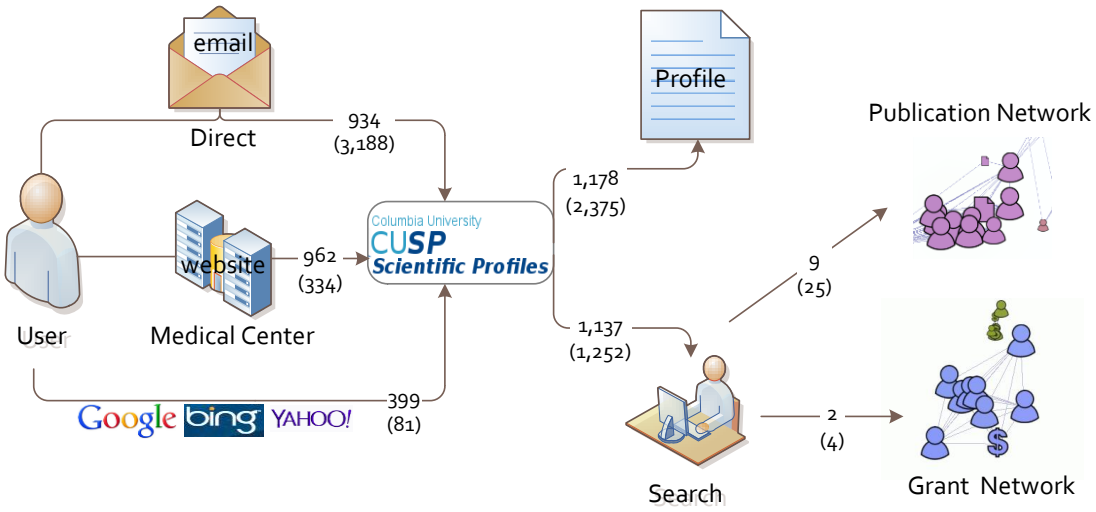


Figure 3. Uses of CUSP in 2013 (numbers outside parentheses) and 2012 (numbers within parentheses)

Discussion

The results of this study have important implications for RNS designs to support biomedical researchers and RNS usage analysis methodology. We observed a slightly increased bounce rate and decreased time spent during the open access mode. However, the small values of RMSE calculated from trends analysis suggest that the temporal change patterns were not much different between the two access modes (**Table 2**). In addition, the average session duration decreased substantially from 11:40 in restricted access to 3:20 in open access. This decrease can be explained by several factors, such as an improved user interface for biomedical researchers in the open access CUSP (**Figure 4**) and different needs of Columbia and non-Columbia users. Future causal studies for these results are warranted.

Results of the trend models developed for this study offer general guidance for RNS design (e.g., open access). The trend models built based on the 100 days of use of each access mode in Figure 2 showed a steady increase of users in open access mode compared to the steady decrease during CUSP's restricted access period. This implies that an open-access model benefits the sustainability of an RNS. The trend models also inform us about how long the effect of marketing (e.g. newsletters) has lasted in Figure 2. Other institutions should consider an open access RNS in order to support interdisciplinary collaboration.

Furthermore, this study presented a novel integration of analytical methods that may be useful to others. The Google Analytics approach provided us with rich time trends data. We were able to obtain log data representing more than 30 different kinds of user behaviors and characteristics. The data were easily imported into various analytical software applications for further analysis, including time series and content mining. While the methods were simple, the results are sophisticated. The time series analyzed for this study have rarely been applied to similar log file studies that use Google Analytics. This trend analysis based on machine learning algorithm is more powerful and sophisticated than the traditional statistical time series analysis using ARMA (autoregressive moving average) or autoregressive integrated moving average (ARIMA) model[7].

In addition, content mining using Automap provides more accurate use information related search terms than the output from Google Analytics. Google Analytics calculates frequency of key terms based on the morphology of the words. In contrast, our content mining approach allows us to aggregate semantically related terms utilizing natural language processing. Google Analytics does not recognize word variation, such as upper vs. lower case, word vs. symbol, singular vs. pleural, and alternate spellings. For example, while Google Analytics considers "comparative effectiveness", "comparative-effectiveness", "Comparative effectiveness" and "Comparative Effectiveness" as four distinct terms, our approach recognizes them as a single concept.

This study inherits the limitations of log file analysis. While this web usage mining provides answers to who, how, where and when questions related to system use, it does not provide answers to why - why users interacted less with certain pages or why users are satisfied or not satisfied with the system. Nevertheless, the usage mining was valuable to assess the global picture of the usage patterns for hypothesis generation to guide further user modeling.

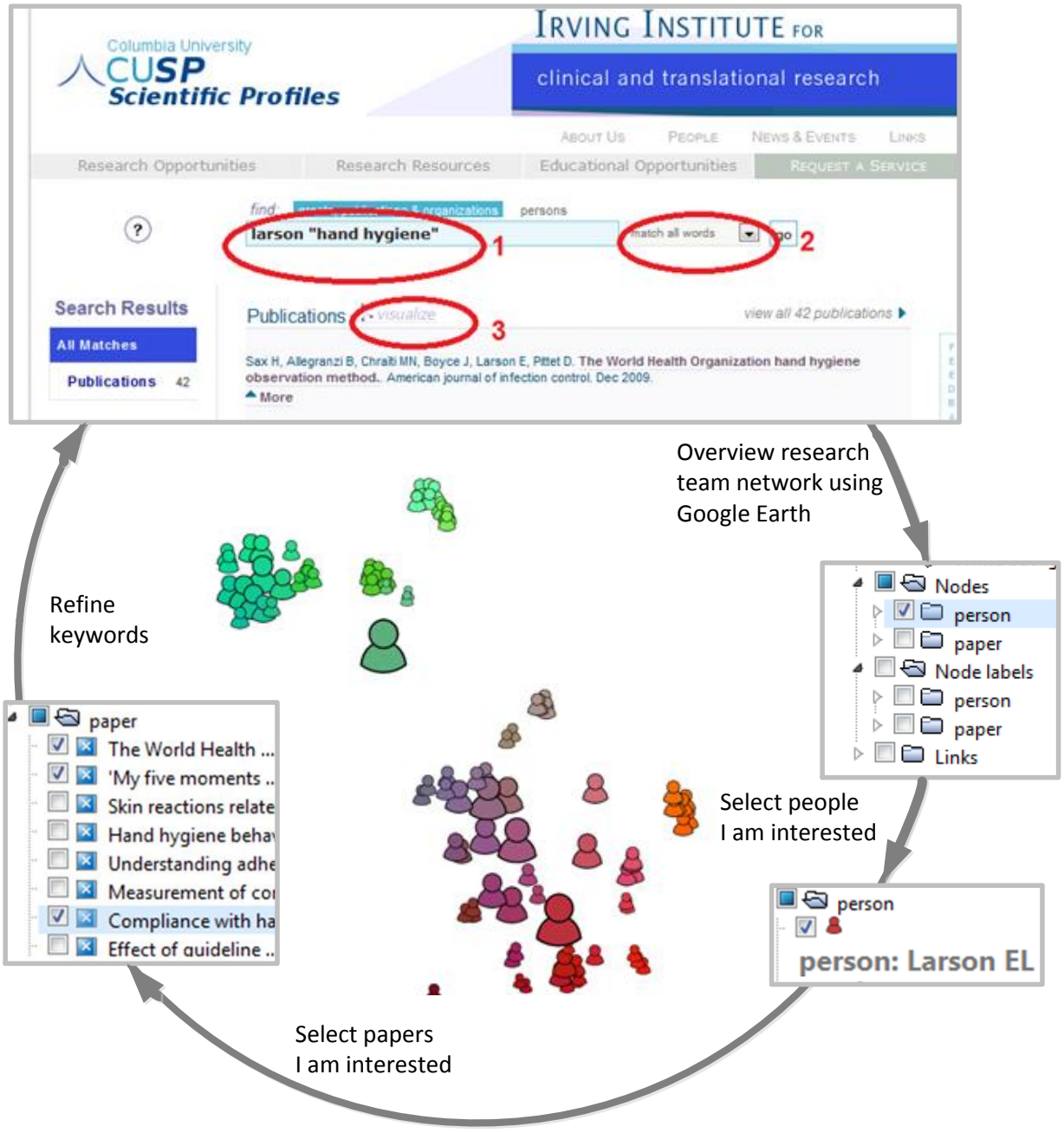


Figure 4. Improved search interface of CUSP and process for biomedical researchers to search for collaborators

Conclusions

This paper reports temporal usage patterns of our locally developed research network system, CUSP, by applying various temporal data mining methods to the Google Analytics usage log. Software used in this study is freely

available on the Internet. The software packages are easy to use by researchers without significant programming skills who need to mine usage patterns on a health-related website. Our temporal usage analytical framework allowed us to efficiently characterize and understand the temporal usage of open-access RNSs like CUSP and our results offer generalized guidance for improving the design of future RNSs.

Acknowledgments

The authors of this study were supported by grants **R01 HS019853** (PI: Bakken), **R01 LM009886** (PI: Weng), and **UL1 TR000040** (PI: Ginsberg). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Liu B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. New York: Springer Heidelberg Dordrecht London; 2011.
2. Boland MR, Trembowelski S, Bakken S, Weng C. An initial log analysis of usage patterns on a research networking system. *Clinical and translational science.* Aug 2012;5(4):340-347.
3. Bracke PJ. Web usage mining at an academic health sciences library: an exploratory study. *J Med Libr Assoc.* Oct 2004;92(4):421-428.
4. Clifton B. *Advanced Web Metrics with Google Analytics.* Indianapolis, Indiana: John Wiley & Sons, Inc; 2012.
5. Bisgaard S, Kulahci M. *Time Series Analysis and Forecasting by Example.* Hoboken, New Jersey: Wiley; 1 edition 2011.
6. Shevade SK, Keerthi SS, Bhattacharyya C, Murthy KRK. Improvements to the SMO algorithm for SVM regression. *Neural Networks, IEEE Transactions on.* 2000;11(5):1188 - 1193.
7. Saigal S, Mehrotra D. Performance comparison of time series data using predictive data mining techniques. *Advances in Information Mining.* 2012;4(1):57-66.

Sample and Clinical Information Link (SCI-Link) - Standardized Data Registry Management System

Monika Ahuja, MCA[1], James Schappet[1], Ryan Lorentzan[1], Yi Wang,MS[1], Ray Hylock,PhD[1], Kimberly K Leslie, MD[2], Heather Davis, MLIS[1], Boyd Knosp, MS[1], Mark Santillan, MD[2], Donna Santillan, PhD[2]

[1]Institute for Clinical and Translational Science, Iowa City, IA; [2]Department of Obstetrics and Gynecology, University of Iowa Hospital and Clinics, Iowa City, IA

Abstract

SCI-Link, a web based application developed by the Institute for Clinical and Translational Science (ICTS), serves as a platform to query EHR as well as sample collection data. This secure integrated platform has been developed to store, curate, manage and share clinical registries in a standard format with multiple research groups collaborating within or outside the University of Iowa. EHR data for patients consented in research studies (as defined in Epic), is extracted periodically and populated in the clinical study defined in SCI-Link. Researchers then have role based access control to subjects in their own studies that they can then query and manage.

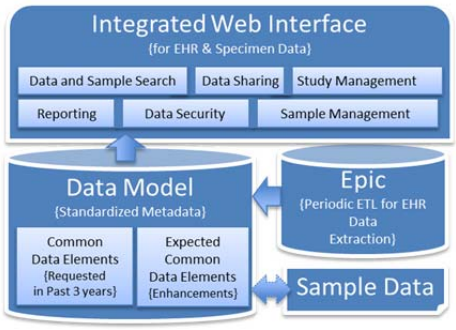
Introduction

Researchers struggle with receiving data from Informatics/IT divisions. Depending on data request queues, and complexity of data requests, data extraction and reporting teams can take anywhere from a few minutes to months to get back to the researchers. Researchers averse to such queues spend valuable time in manual EHR data accumulation and management. Our novel web based application SCI-Link automates data extraction of commonly requested data elements in standardized study-specific data registries. With customization of ETL (Extract Transform Load), this resource can be used by other biobanks.

Methods

The OBGYN team started their IRB-approved study with a small set of patients. Data was initially manually extracted from Epic and managed along with sample collection data into Excel files. As the patient population increased, data collection and management efforts increased significantly and made it difficult to locate/manage samples and associated clinical data. The researchers collaborated with ICTS for development of a system that could migrate their specimen data from Excel files, integrate it with the automatically extracted EHR data which could then be queried. As a result SCI-Link was designed and developed with required ETL, data security, data sharing, and data search features. OBGYN team maintains consented patient lists in Epic, ETL is performed periodically to extract and load data for these patients into a generalized data model and users are allowed to capture/manage the study /authentication/other administrative data along with the specimen collection data. Built in query features allow users to search available samples and/or EHR data based on diverse clinical conditions of patients.

Implementation



Results

This tool has saved significant time and effort for this team in terms of data extraction, management, security, sharing, and storing the sample and EHR data. Search capability has enhanced the capacity of this team to collaborate on diverse research questions because the clinical and specimen information is readily available in *one* secured application that can be easily accessed/queried by all research team members.

Conclusion

SCI-Link helps researchers by saving time in manual data accumulation, preventing need for proof-reading data entry and manual de-identification of data and submission of multiple data requests to IT team. In addition, the main benefit is that this tool links the sample and clinical data allowing for rapid identification of samples for research use. Furthermore, this program benefits the ICTS Informatics team because it ensures that data registries consist of common data elements that are consistent across different studies, provides a single platform for the institution which creates the opportunity for organizational-level view of data and ensures that study data is extracted from an authoritative source, and, most importantly, that data is managed in a centralized secured location and not circulating in diverse formats and unsecured forms (like Excel files/distributed databases).

This publication was made possible by Grant Number 2 UL1 TR000442-06 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research.

A Decision Framework for Selecting a Federated Data Sharing Platform

Michael J. Ames[1], MBI. Jessica Bondy[2], MSHA. Ted D. Wade[3], PhD. Arthur Davidson[4], MD, MSPH. Michael G. Kahn[2], MD, PhD.
[1]University of Colorado, Aurora, CO. [2]Colorado Clinical and Translational Sciences Institute, Aurora, CO. [3]National Jewish Health, Denver, CO. [4]Denver Public Health, Denver, CO.

Summary: *Institutions need to share information without compromising control over their own data resources. This need has driven a proliferation in technical platforms for federated data sharing. However, no single platform has yet emerged that satisfies every use case. We present a decision tool to help institutions determine which platform is best suited for their federated data-sharing needs.*

Introduction and Background. The Colorado Health Observations Regional Data Service (CHORDS) is a collaboration between several affiliated but independent research and health care organizations in Colorado. Its purpose is to provide the technical and policy infrastructure for regional data sharing to support cohort discovery, cost/quality research initiatives, regional health surveillance, and targeted primary care interventions in public health.

We previously presented an analysis of several federated data sharing platforms, identifying an optimal solution for our specific CHORDS requirements[1]. In the course of this project it became apparent that there was no "one-size-fits-all" solution. Each platform fulfills some requirements but not others. We will present a decision tool developed by CHORDS to aid in identifying the appropriate platform based on key differentiating requirements.

Methods. The federated data sharing landscape was surveyed to develop a list of candidate systems to include in the analysis. Survey sources included the Query Health Summer Concert Series, a literature search, and the knowledge of CHORDS team members. We focused on candidate systems with free availability and broad community adoption. The capabilities of the candidate systems were analyzed with respect to several use cases, and a capability matrix was developed to serve as the basis for the proposed decision tool. Current candidate systems include i2b2/SHRINE, TRIAD, and PopMedNet.

Results. The results section will consist of the decision tool itself, which is still under development, as the centerpiece of the presentation. It will provide the community with a simplified approach to matching key requirements against the capabilities of the platforms analyzed. For example, we anticipate that the following questions will be among the key differentiators: *Do you need data sharing institutions to manually review and approve queries before releasing data? Do you need real-time results from queries? Do you have in-house technical expertise in Java? Is this for research or public health surveillance? What are your major regulatory requirements?* Questions like these will be incorporated into a flowchart or decision tree.

Discussion. We will review the lessons learned from the process of developing the decision tool, including recognition that the evolving nature of federated data sharing technology, may require the tool to be periodically updated. Some platforms will have necessarily have been omitted from the analysis, such as commercial platforms, and will be acknowledged. In addition, we will discuss weaknesses observed in all federated data sharing platforms, and suggest some ways in which they should be improved in order to bring greater benefit to the research informatics community.

Conclusion. The need for institutions to share data without compromising control over their own data resources will continue. At present, federated data sharing is the most promising approach to meeting this need. However, there is no universal solution available today. Greater transparency about the capabilities of each platform and their ideal uses, in addition to tools such as the one presented here, will help institutions make better informed decisions about platform selection.

References

1. Ames MA et al (2012, March). *Analysis of Federated Data Sharing Platforms for a Regional Data Sharing Network.* Poster presented at AMIA CRI, San Francisco, CA.
2. Davidson A et al (2013, June). *Scaling a Multi-purpose Distributed Registry Network*. Poster presented at Academy Health, Baltimore, MD.

Weaving a Strong Trust Fabric through Community-Engaged Research: Lessons from the WICER Project about Digital Infrastructure for the Learning Health System

Suzanne Bakken, PhD, RN, FAAN, FACMI[1,2], Niurka Suero-Tejeda, MS, MA, CHES[2], J. Thomas Bigger, MD[3], Adam Wilcox, PhD, FACMI[4], Bernadette Boden-Albala, DrPH[5]
[1]Department of Biomedical Informatics, [2]School of Nursing, [3]Department of Medicine, Columbia University, New York, NY, [4]Intermountain Health Care, Salt Lake City, UT, [5]ICAHN School of Medicine, Mount Sinai Medical Center, New York, NY

Introduction and Background: An Institute of Medicine report identified 12 characteristics of the Learning Health System. Among these is the need to build a strong fabric of trust among stakeholders through communication and demonstration of value. A follow-up workshop endorsed three principles: 1) build a shared learning environment; 2) engage health and health care, population and patient; and 3) leverage existing programs and policies. The difficulty of building a strong fabric of trust among racial and ethnic minorities has long been acknowledged and is evidenced in the literature by low participation rates in research studies and biobanks and limited use of information technologies for health-related purposes. The objective of this presentation is to share the processes that we implemented to foster a strong trust fabric in the community component of the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project and the resulting impact.

Methods: We applied multiple community engagement approaches during the process of collecting survey data from community residents including: building upon the established collaborations of the CTSA-funded Columbia-Community Partnership for Health (CCPH); free community blood pressure screening and education at CCPH; focus groups to inform survey content; data collection by bilingual community health workers from Washington Heights/Inwood in homes, community organizations, and local businesses; incorporation of snowball sampling methods; and compensation for participant time with incentives of value to residents (e.g., grocery coupons). As part of the informed consent process, we also queried individuals regarding their preferences for linking survey data with clinical data, being contacted for future research studies, willingness to provide biospecimens.

Results and Discussion: Almost 90% survey participants agreed to linkage of survey and clinical data. The great majority also indicated their willingness to be contacted for potential participation in future research studies by the WICER team or other investigators. We were successful in recruiting participants for ancillary WICER studies such as focus groups to test visualizations of survey data. Moreover, investigators outside of WICER who contacted WICER participants meeting their inclusion criteria for other studies reported high percentages of enrollment. We achieved our project goals of collecting dried spots (n=500) and saliva for hormone assays (n=250) in a relatively short period of time – 7 weeks for the latter as part of follow-up data collection for the survey at CCPH. We exceeded our initial goal of 1,000 saliva samples for DNA analysis collecting more than 1,600 samples. Not all participants were approached given that biospecimen collection started after the first 2,000 surveys were collected so there is potential for that number to increase. Although we can only compare our results to the literature rather than to baseline in our community, we believe that our findings provide evidence of an emerging fabric of trust in the Washington Heights/Inwood community and contend that it is the result of our community-engaged approaches. We are continuing to build this trust fabric by returning data to survey participants, community-based organizations, and the community at large through a digital infrastructure that includes access to infographics tailored to the needs of the stakeholders. This infrastructure creates a foundation for self-management and community-level health promotion strategies.

Acknowledgments: R01HS019853, R01HS022961, NYS Department of Economic Development NYSTAR (C090157)

Utilizing temporal information to improve adverse drug event prediction models

Aurel Cami[1,2], Ben Y. Reis[1,2]
[1]Division of Emergency Medicine, Boston Children's Hospital, Boston, MA, USA
[2]Department of Pediatrics, Harvard Medical School, Boston, MA, USA

Abstract
This study focuses on leveraging temporal changes in adverse drug event data to improve adverse event prediction.

Introduction
Predictive models of unknown adverse drug events (ADEs) can be useful for reducing the vast space of possible associations among drugs and ADEs. These models are evaluated by comparing their predictions with newly discovered drug-ADE associations. Since newly discovered associations constitute only a small percentage of all possible drug-ADE pairs, a model may exhibit a low positive predictive value (PPV) even when it has high sensitivity and specificity. A model is most practical when it generates a relatively small set of high-value targets for further investigation, so increasing the PPV is essential to improving the practical value of these models. Here, we investigate whether features extracted from recent changes in drug-ADE data over time could be used to increase the PPV of Predictive Pharmacosafety Networks (PPN) models [1].

Methods
We analyzed six snapshots of the Lexicomp clinical drug safety database taken annually between 2005 and 2010. We mapped the ADE names contained in all six data sets to the High-Level Terms (HLTs) of the Medical Dictionary for Regulatory Activities (MedDRA) standard. We used the 2007 data set – i.e. all historical data accumulated till 2007 – to train a PPN model (*Historical Model*) and to generate predicted scores $S_{historical}$ for all drug-ADE pairs that were not reported as associations through 2007. Next, we fit a second model (*Recent Change Model*) based on the recent changes in the drug-ADE data, i.e. associations that were newly reported in 2006 or 2007. The explanatory variables for this model included PPN covariates derived from a network containing only the newly reported 2006 and 2007 associations, as well as the top-level Anatomical Therapeutic Chemical (ATC) code of drug, top-level MedDRA category of ADE (SOC), and the number of years the drug had been on the market. Based on this model, we generated a second set of scores $S_{recent\text{-}change}$. We computed predictive performance metrics for both models using the 2010 data as a validation set. Finally, we compared the PPV computed from $S_{historical}$ scores with the PPV of a third model (*Hybrid Model*) that multiplicatively integrated both scores based on historical and recent-change data.

Results
Among 762 drugs and 866 ADEs, a total of 35,772 associations were reported in the 2005 data set, 6,113 new associations were reported in the 2006-7 data sets, and 5,358 further new associations were reported in the 2008-10 data sets. Comparing predictions with the 2010 data, the *Historical Model* achieved an area under the ROC curve (AUROC) of 0.87, while the *Recent Change Model* achieved an AUROC of 0.67. The *Hybrid Model* achieved markedly improved performance for specificity values higher than 98.5% – corresponding to practically useful predicted sets with up to 800 true positives (Figure 1A) – with relative increases in PPV as large as 49% (Figure 1B).

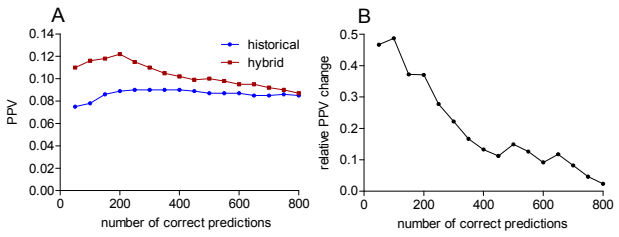


Figure 1. (A) PPV corresponding to historical and hybrid scores; (B) Relative change in PPV.

Conclusion
This study indicates that features extracted from recent changes in drug-ADE data can be used to improve the predictive performance of PPN models. We found that combining a model trained on recent changes in the data with one trained on complete historical data, leads to increased positive predictive value for high-specificity settings representing predicted sets with up to hundreds of true positives.

References

1. Cami A, Arnold A, Manzi S, Reis B. Predicting adverse drug events using pharmacological network models. Sci Tr Med, 3(114), 114ra127, 2011.

National Institutes of Health's Biomedical Translational Research Information System (BTRIS) Data Query Tool

James J. Cimino, MD, Elaine J. Ayres, MS, RD
Laboratory for Informatics Development
NIH Clinical Center
Bethesda, MD

The National Institutes of Health's Biomedical Translational Research Information System (BTRIS) is a repository of intramural clinical research data collected from 1976 to present, in two NIH electronic health records and a variety of clinical trials data management systems.[1] In addition to providing access to identified data from active clinical studies to the investigators on those studies, BTRIS also provides access to data without personal identifiers from across all studies to intramural researchers. A preliminary version of this latter function was created as proof of concept and generated little interest. At the same time, researchers were requesting de-identified data that required more sophisticated, manually mediated queries. A new query interface was designed based on the requirements for those more complex queries and the ability of the new interface to service those queries has been reported.[2]

The BTRIS repository is a Microsoft SQL-server database that partitions data into "events" and "observations" and then further partitions data into "measureable" (laboratory and vital signs-related), "substance" (medication-related) and "general (other, including text reports). Each data domain (demographics, laboratory, medications, problem lists, etc.) are assigned to one particular set of event and observation tables. All data are in identified form, but most fields within an observation or event are free of personally identifiable information. The user interface was modeled on the i2b2 Workbench design[3] using NET, HTML5, JavaScript/jQuery, XML and JSON.

From March 1, 2013 to present, the system has been used 957 times to produce summary results, with downloading of detailed data sets 38 ranging in size from a few hundred rows of data to over one million rows. The presentation will provide a brief tour of system features,[4] summarize the first year's experience with usage, describe efforts to remove identifiers from clinical text, and list the steps necessary to interface the query tool with other databases.

1. The National Institutes of Health's Biomedical Translational Research Information System (BTRIS) is a repository of intramural clinical research data collected from 1976 to present, in two NIH electronic health records and a variety of clinical trials data management systems.[1]
2. Cimino JJ, Ayres EJ, Beri A, Freedman R, Oberholtzer E, Rath S. Developing a self-service query interface for re-using de-identified electronic health record data. *Stud Health Technol Inform*. 2013;192:632-6.
3. Murphy SN, Mendis M, Hackett K, et al. Architecture of the open-source clinical research chart from Informatics for Integrating Biology and the Bedside. AMIA Annu Symp Proc. 2007 Oct 11:548-52.
4. http://www.youtube.com/watch?v=5yHiVSMew7k

Who's Counting? Probability Statements and Risk in Medical Literature

Léa A. Deleris, PhD[1], Lamia Tounsi, PhD[2], Bogdan Sacaleanu, PhD[1],
[1]IBM Research, Dublin, Ireland; [2]Dublin City University, Dublin, Ireland

Abstract

This paper investigates how to extract quantitative risk information - in the form of probability statements - from academic medical papers. Our approach is based on regular expression matching of probability numbers followed by extraction of the associated variables through machine learning using a combination of lexical, syntactic and semantic features. For that purpose, we have first developed a reference corpus which is composed of Medline abstracts related to risk factors for breast cancer. Our preliminary results, based on the detection performance metric, are encouraging. We obtained a 92% F-value for probability number detection and between 50% and 75% for variable identification depending on the type of variable. However, those results are typically associated with higher recall than precision. Research improvements and extensions are thus warranted before being able to fully make use of probability statements extracted from a discovery support system.

Introduction

Medline provides a sizeable amount of medical information on the web in the form of research papers and abstracts which has been steadily growing at a rate of about 1M publications per year in the past few years. Manual consumption of this information is no longer practical and researchers have turned to automated processing for instance for literature-based discovery[1,2]. In this research, our long-term objective is to build algorithms to extract risk information from medical papers. Such information includes (i) probability statements indicating for instance the likelihood of occurrence of a disease and (ii) dependence information indicating influence among risk factors (such as smoking on lung cancer). The output knowledge base could then provide health experts with a synthetic view of risk factors associated with a disease and/or diseases associated with a risk factor. Also, it would retain links to the individual sources of information, thus providing detailed provenance information and enabling human feedback. In addition, it could be aggregated to highlight consensual opinions and disagreements and provide a mean for monitoring factual trends over time, which would be useful for clinical research.

We focus in this paper on the extraction of probability statements as a richer body of work exists in the domain of risk-factor and causal relations extraction[3,4]. A well-formed probability statement, $P(A \mid Z) = x$, is associated with two sets of variables: A representing the main variables and Z the conditioning variables (which may be empty). In our case we make a distinction in the set Z between the main set of conditioning variables B and the context variables C. For instance, if A represents breast cancer incidence, B could represent whether or not women have received hormone replacement therapy and C would indicate the population sub-groups to whom the statement applies (based on the study settings), for instance Irish women aged 50-60 year old. While artificial from a probability perspective, the distinction between sets B and C is helpful for the annotation task and the recovery of the variables, especially for probability statements in the form of odd ratios and relative risk. In fact, in addition to simple probability statements, we seek to extract other variations in the form of odds, odds ratio, hazard ratio and relative risk as they are frequent in the medical literature.

Consider the sentence: "*Average duration of breast-feeding of 11-12 months reduced risk of breast cancer by 54% compared with the duration of 1-4 months (odds ratio 0.46; 95% confidence interval, 0.30-0.70)*"[PMID 17297388]. We seek to extract the following structured information:

- A: *risk of breast cancer*
- *B: Average duration of breastfeeding of 11-12 months*
- $\overline{B}$ *(complement of B): the duration of 1-4 months*
- $P(A \mid B) / P(A \mid \overline{B})$ =1-**54%**(= 0.46) (relative risk)
- $[P(A \mid B) / P(\overline{A} \mid B)] / P(A \mid \overline{B}) / P(\overline{A} \mid \overline{B})$ = **0.46**

We are in the initial phases of this research endeavor. To this point, our focus has been on identifying the relevant elements (probability numbers and variables). We will address the classification of probability numbers into subtypes and the reconstruction of the statements in our next steps.

Background

Supervised machine learning and Natural Language Processing (NLP) methods have been used successfully for relation extraction, entity end event detection in a variety of domains including biomedical (BioNLP shared tasks[5,6]) and clinical data (i2b2 shared tasks[7,8]). In this paper the task is to identify the segments of text representing both probability numbers and variables associated with probability statements. While similar to medical entity recognition[8], our task is more complex as the variables being considered are more diverse than those considered in previous medical entity extraction works; they can range from diseases to symptoms, treatments, genes, proteins, behaviors or any medical entity of interest. Different entity types usually call for different approaches that either leverage morphosyntactic and orthographic properties of entity words[9] (which vary across types) or assume an internal structure of an entity like a noun phrase[10]. Therefore in devising an overarching method for all possible types, one has to abstract away from this level of detail or assumptions. Accordingly, our proposed approach treats entity detection as a classification problem with as few lexical features as possible and considers each individual token as a potential candidate to a variable content.

To the best of our knowledge, very few research papers tackle the specific problem of extracting numerical probability statements. Some efforts have been spent in inferring probabilities based on co-occurrences counting PubMed abstracts[11,12]. By contrast, we seek to extract explicit probability statements. In that sense, research focused on the extraction of structured information from reports about randomized control trials[13] is closer to our objective given the need in that context to also retrieve hazard rate and relative risk information.

Data

We created a specific corpus which consists of a collection of 200 abstracts from MEDLINE selected according to the query: "KW: breast cancer AND parity" over the past 5 years. ProbMed-200 has been annotated using a web-based tool for text annotation BRAT (http://brat.nlplab.org/). The annotation of 50 abstracts was performed by two researchers (one risk expert and one NLP expert) and one undergraduate NLP student independently and then reconciled manually through discussions. The annotation of the remaining 150 abstracts has been completed by the student only. The annotation of probability numbers shows a strong agreement among annotators (kappa scores 0.89-0.92), while the scores for the variable annotation task are lower (0.76-0.82 for variable A and 0.73-0.8 for variable B). Our investigation revealed that often the boundaries of the same labeled instance vary across different annotators, e.g. from "*Malaysian women have 1 in 20 chance of developing breast cancer in their lifetime*", annotator 1 (resp. 2, resp. 3) labels "breast cancer" as variable A (resp. "developing breast cancer in their lifetime" resp. "breast cancer in their lifetime"). For this reason we use the detection performance (DP) as evaluation metric to report all the scores in this paper. DP is a loose measure of correctness[14], which is generous in the sense that it considers any system output correct as long as at least one token overlaps with the reference expression. Therefore, the performance results reported here are optimistic as they assume that through post-processing we are able to recover a variable from one token. From the annotated data we created three data sets as presented in Table 1. The training set is used to train the classifier, the validation set to determine the best sets of parameters and features for each classifier and the test set for an unbiased evaluation of the classifiers (using best parameters and feature sets).

Table 1. Characteristics of the Training, Validation and Test Sets.

	Train	Validation	Test
# of Sentences	897	514	433
# of Sentences containing a probability number	212	82	78
# of Probability Numbers	531	205	219
# of Variables A	254	114	85
# of Variables B	276	99	114
# of Variables C	89	33	31

Method & Results

Our process to extract the elements of a probability statement follows two steps: First we identify the tokens of a sentence that correspond to a probability number. Second we use that knowledge to detect and categorize variables.

A. Extracting Probability Numbers

Probability numbers *x* are numerical expressions which can be represented by (i) digits e.g. *"20%",* (ii) multi-word sequences e.g. "*twenty percent*", *"one in twenty"*, (iii) mix of both e.g. "*20 percent*", "*1 in 20*".

We manually defined a set of heuristics to search for regular patterns in the text representing a probability value using the instances present in the training set e.g. [0-9]+%, [0-9]+ in [0-9]+, OR= [0-9]+.[0-9]+[0-9]+, etc. We also investigated machine learning approaches using our training set. All sentences were tokenized and parsed and then provided to both a support vector machine (SVM) and Naive Bayes classifier built using combinations of the following features: token, frequency, part-of-speech tag assigned by a statistical parser[15], lexical patterns such as whether the token is a digit sequences, a percentage, "OR" (Odds Ratio), "RR" (Relative Risk) and syntactic information such as the type of the phrase containing the token and if the token is head of this phrase or appears on the left side or the right side of the head. The F-value for the rule-based approach is 92% while SVM's performance was 53.5% and Naïve Bayes only 42% when using all features. In fact, the scores obtained with different subsets of features were lower for both algorithms. At this point in our research, we feel that the rule-based approach performs well enough for not investigating further what kernel or additional features would enable the classifiers to improve and outperform the rule-based approach. However, such exploration is left for future research.

B. Characterizing Associated Variables

To identify the variables associated with probability numbers, we consider 26 linguistic features covering the lexical, syntactic, and semantic aspects of the data. The full list is presented below and discussed thereafter.

Lexical

- ***Type:*** Part-Of-Speech tag, Out-Of-Vocabulary
- ***Orthographic:*** Index, Capital, Digit
- ***Probability-related :***Probability number, Minimum distance to probability number

Syntactic

- ***Constituency trees :*** Phrase tag, Phrase length, Head/Left/Right of the phrase, Phrase tag of the ancestor head, Same predicate as probability number, Distance to predicate, Constituency path contains clause tag
- ***Dependency trees:*** Dependency tag, Dependency tag of dependent, Dependency tag of the head of the phrase, Distance to predicate, Dependency path contains subject, Dependency path contains object

Semantic

- ***ULMS***: Semantic group, Group appears in most frequent

Variables from our training set tend to be either noun phrases (NP) or prepositional phrases (PP) with different levels of complexity. Variable A, which captures the risk, is likely to be a concise segment of text in the form of an NP phrase. When it shares the same predicate verb with a probability number it is often an object of this verb. Variable B, which describes the risk factors, is likely to be a long segment of text in the form of a complex NP with embedded PP phrases. When it shares the same predicate verb with a probability number it is often a subject of this verb. In general also the position of Variable A (resp. B) in the text is often close (resp. distant) to the probability number. Variable C, which provides contextual information, is less frequent. It usually consists in a long segment in the form of a PP phrase which appears at the beginning or end of a sentence. These observations led us to include Part-Of-Speech (POS), Index (location of the token in the sentence) and probability related characteristics (indicator of probability number and minimal distance to a probability number) in our lexical feature set. We also identify tokens that are considered as Out-Of–Vocabulary (i.e. not seen often in the training set) as most of abbreviations and proper nouns appearing in the test set are unknown in the training set.

Similarly, in the syntactic domain, our choice of feature is driven by our analyses of the annotated set. Constituency features include the phrase tag of the parent node as variables are more likely nominal phrases. We compute the length of the phrase and whether the current token is the head or appears to the left or right of the head to make use of the fixed-word order of the English language. We look at whether the token shares the same predicate with the probability number as some variables do, especially Variable A (even in complex and long sentences). To provide information about length and complexity of the internal structure of the trees, we include the number of steps in the constituency path to predicate and whether the path contains a clause tag (such as S, SBAR, etc). In fact we observed that the trees representing Variable B are larger and more complex than those representing Variable A. We also check whether this path contains the phrase tag of the head ancestor. This feature provides information about the size of the span representing the variables. Dependency features are very much translations of the constituency features in the dependency tree (whenever relevant).

In parallel, we analyzed the semantic characteristics of variables using the Interactive MetaMap service. We define semantic patterns using the most frequent semantic groups representing the variables with a frequency threshold set to 5 instances[16]. Variables A covered 4 semantic types (Concepts & Ideas, Disorders, Anatomy, Physiology),

Variables B covered 6 types (Concepts & Ideas, Living Beings, Disorders, Physiology, Procedures, Chemicals & Drugs) and variable C covered 5 types (Concepts & Ideas, Disorders, Physiology, Living Beings, Activities & Behaviors). As there is some overlap of concepts among the variables, the semantic group feature may only provide a weak signal as to whether and of which kind of variable a token belongs. In addition to the semantic group type, we also added a feature indicating whether this group is associated with the most frequent semantic groups representing Variable A (resp. Variable B and C).

C. Extracting Associated Variables

We investigated five machine learning methods to label the variables: Naive Bayes (NB), Logistic Regression (LR), K-nearest neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT). Given variables can represent many types (diseases, symptoms, genes, treatments), we focused in this initial stage on exploring a varied range of approaches. We did not investigate conditional random fields as we felt the multiplicity of types would make context less relevant, yet we will consider this algorithm in our subsequent analyses. For each algorithm, we optimized its parameters based on the performance on the validation set and also performed greedy forward feature selection. Depending on algorithm and variable, the number of features finally used varied from 1 to 23. Typically, Variable A requires less features (average of 7 features) than the other two (average of 13 features) and SVM results in the sparsest sets overall (Variable A: 2, Variable B: 5 and Variable C:1). In our experiments, we assume that the data has already been perfectly filtered with sentences containing probability numbers. Recall also that all metrics provided in the table are established based on the detection performance metric.

Table 2. Performance on Variable Identification.

Variable	Metrics	KNN	NB	DT	SVM	LR
Variable A	F-value	0.49	**0.57**	0.50	0.48	0.48
	Precision	0.39	0.41	**0.43**	0.32	0.37
	Recall	0.65	0.93	0.61	**0.97**	0.69
Variable B	F-value	**0.75**	0.69	0.61	0.36	0.70
	Precision	**0.79**	0.65	0.61	0.47	0.68
	Recall	**0.72**	0.72	0.61	0.29	**0.73**
Variable C	F-value	0.49	0.49	**0.51**	0.32	0.45
	Precision	**0.74**	0.54	0.54	0.23	0.36
	Recall	0.36	0.45	0.48	0.55	**0.58**

For Variable A, Naïve Bayes out-performs the other algorithms as can be seen from the F-value. In fact the Naïve Bayes algorithm presents the most sensible balance of precision (0.41 against 0.43 for the best algorithm) and recall (0.93 against 0.97 for the best algorithm). In our application, we care more about precision than recall, as missed instances can be compensated by volume and the fact that medical papers tend to be redundant while spurious instances will results in frustration on the human expert side in the short term and loss of confidence in the information retrieved in the long-term. For Variable B, performance is generally higher than for Variable A and KNN is the best algorithm, dominating almost all others on all three criteria. There could be one simple mechanical explanation: Variables B tend to be longer than Variables A and the detection performance metric will favor longer variables, everything else being equal. Finally for Variable C, all algorithms perform relatively poorly. For this variable, the trade-off precision recall is more pronounced.

D. A Few Examples

To provide additional insights into the current status of our approach, we present here a few illustrative sentences along with the associated output. For each sentence, we highlight the variables that should be identified, indicate their types and the number of classifiers, out of the 5 considered, that correctly detected them, even if only partially.

"In contrast, there was no association with ***current smoking*** $_{(VarB-3)}$ *and* ***breast cancer death*** $_{(VarA-5)}$ ***;*** *the RR (95 % CI) was 1 (0.83-1.19)."* [PMID 17278091]

"Overall ***SNCG mRNA expression*** $_{(VarA-5)}$ *was detectable in 36 %* ***breast cancers*** $_{(VarB-2)}$*."* [PMID 16821081]

*"****Chinese women*** $_{(VarB-5)}$ *were twice more likely than* ***Malay women*** $_{(VarB-4)}$ *to* ***have a mammogram done*** $_{(VarA-3)}$*."* [PMID 17245514]

*"**Among post-menopausal Hispanic women recently exposed to hormones**$_{(VarC-5)}$ **the A allele of the -202 C > A IGFBP3 polymorphism**$_{(VarB-5)}$ increased risk of **breast cancer**$_{(VarA-4)}$ (OR 1.5 , 95 % CI 1.06-2.33)."* [PMID 17051426]

All the sentences above are adequately processed. Note that they are all fairly short, straightforward in their structure and containing only one probability number, though it can be in the form of text ("*twice*"). By contrast, the sentences presented below are much more challenging.

*"Only 4 % of **basal-like carcinomas**$_{(VarB-1)}$ showed **MYC amplification**$_{(VarA-4)}$, compared to 8.75 % and 10.7 % of **lumina**$_{(VarB-0)}$ **l** and **HER2 tumours**$_{(VarB-1)}$ respectively."* [PMID 17158641]

No algorithm completely identified the variables but at least one of them partially identified the first and third Variable B, which implies that they can be recovered through post-processing. However, the second Variable B has been missed by all algorithms.

*"**Breastfeeding**$_{(VarB-2)}$ was associated with **a reduced risk**$_{(VarA-3)}$ for **carriers of BRCA1 mutations**$_{(VarC-0)}$ (0.74 [0.56-0.97] ; p=0.03)."* [PMID 17196508]

In the above sentence, Variable C is missed altogether which can be misleading as it represents an essential distinction in the associated paper which investigates the difference between BRCA1 and BRAC2 mutations carriers and the risk of breast cancer. In addition, the extracted Variable A is *"reduced risk"* which means that anaphora resolution needs to be done to clarify what it is referring to. The sentence below is an even more salient example of this issue as Variable A is only implied.

*"The association with **tubal ligation**$_{(VarB-2)}$ was not significant for **carriers of BRCA1 mutations**$_{(VarC-1)}$ (0.8 [0.59-1.08]; p=0.15), or for **carriers of BRCA2 mutations**$_{(VarC-0)}$ (0.63 [0.34-1.15]; p=0.13)."* [PMID 17196508]

Discussion and Conclusion

The long-term goal of our research is to build algorithms to automatically extract probability statements from medical journal and transpose them into a structured data model that would serve as the basis of a discovery support system. This paper presents results associated with extracting probability numbers and identifying the associated variables which we undertake using a mixture of rule-based and machine learning approaches. Even in this intermediary form, extraction results are informative for clinical research: by providing the source sentence side by side with the structured information extracted, a human, (with enough knowledge in the domain) is able to determine right away the type of probability number that is mentioned.

Performance for identifying probability numbers is above 90% while locating the associated variables proves more challenging, with an F-score which varies from 75% for variable B (representing risk factors) to 57% for variable A (representing risk) down to 49% for Variable C (representing the context). We believe that the findings discussed in this paper are encouraging in that they confirm the feasibility of our extraction tasks. However, the results on variable extraction are typically associated with a relative high recall and a mediocre precision, while for practical purpose we care more about the latter. We therefore need to improve the proposed approach to increase precision while not diminishing recall too much.

We have identified several directions for such improvements. First we observed that some algorithms were much more efficient on specific types of probability numbers, for instance, for Variable C, KNN is very effective when the probability number corresponds to a simple probability statement. Consequently we may modify our process by first classifying probability numbers into types and then developing different variable extraction algorithm for each type. In the same perspective, we have seen that simple sentences with only one probability number were better processed than complex ones. We will therefore explore whether creating categories of sentences based on characteristics such as number of probability numbers, length, presence of key terms such as "*compared to*" could be helpful to better guide the variable extraction process. In addition, to reduce the number of spurious variables recovered, we will investigate multi-class classification rather than parallel binary classifications as we have done here.

Beyond addressing the limitations of the results presented in this paper, a key next step to make this research practically meaningful is to be able to reconstruct the probability statements by associating variables to each number. We also plan on expanding to verbal description of risk (e.g., impossible, unlikely, probable) even though they tend to be fairly uninformative in the sense that the mapping to actual numbers can be a wide interval[17].

References

1. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7–18.
2. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting Semantic Relations for Literature-Based Discovery. AMIA Annu Symp Proc. 2006:349-53.
3. Hamon T, Graa M, Raggio V, Grabar N, Naya H. Identification of relations between risk factors and their pathologies or health conditions by mining scientific literature. Stud Health Technol Inform. 2010;964-8.
4. Blanco E, Castell N, Moldovan D. Causal relation extraction. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 2008.
5. Kim JD, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 Shared Task on Event Extraction. Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, 2009: 1-9.
6. Kim JD, Pyysalo S, Ohta T, Bossy R, Nguyen N, Tsujii J. Overview of BioNLP Shared Task 2011. Proceedings of BioNLP 2011 Workshop, 2011:1-6.
7. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. Journal of the American Medical Informatics Association. 2010;17:514-518.
8. Uzuner O, South BR, Shen S, DuVall SL. i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 2011;18(5):552–556.
9. Abach AB, Zweigenbaum P. Medical Entity Recognition: A Comparison of Semantic and Statistical Methods. Proceedings of BioNLP 2011 Workshop, 2011;56–64.
10. Frunza O, Inkpen D. Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences. In Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. 2010;91-98.
11. Theobald M, Shah N, Shrager J. Extraction of conditional probabilities of the relationships between drugs, diseases, and genes from PubMed guided by relationships in PharmGKB. Summit on Translat Bioinforma. 2009;2009:124-8
12. Sanchez-Graillet O, Poesio M. Acquiring Bayesian networks from text. Proceedings of Fourth International Language Resources and Evaluation (LREC'04). 2004.
13. Hsu W, Speier R, Taira K. An automated pipeline that extracts information related to the statistical analysis from full-text RCT literature, mapping this information to a logical data model. AMIA Annu Symp Proc. 2012; 2012: 350–359.
14. Olsson F, Eriksson G, Franzén K., Asker L, Lidn P. Notions of Correctness when Evaluating Protein Name Taggers. Proceedings of the 19th International Conference on Computational Linguistic (COLING) 2002;1
15. Attia M, Foster J, Hogan D, Le Roux J, Tounsi L and van Genabith J.. Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French. Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages , 2010;67-75,
16. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003;36(6):414-432.
17. Hanauer DA, Liu Y, Mei Q, Manion FJ, Balis UJ, Zheng K. Hedging their Mets: The Use of Uncertainty Terms in Clinical Documents and its Potential Implications when Sharing the Documents with Patients. AMIA Annu Symp Proc. 2012;2012:321-30.

Leveraging Big Data Technology within i2b2 Platform

Xiao Dong PhD, Neil Bahroos MS, Morris Chukhman MS Eugene Sadhu MD, Robert Johnson BS, Himanshu Sharma MS, Denise M. Hynes, PhD, RN University of Illinois at Chicago, IL 60612

Abstract

In this work, we present our experiences incorporating Big Data technology, specifically Apache Hive, into the i2b2 framework, and illustrate the strengths of combining these two frameworks. Our approach has two main advantages: (a) the same i2b2 cohort identification querying tools can be executed against the Hive Database with minimal modification to i2b2's CRC cell; (b) leveraging the readily available parallelism of Apache Hadoop without having to implement MapReduce jobs. We further demonstrate that such a hybrid approach provides a scalable solution to enhance the existing capability of i2b2 to meet the rapidly increasing "Big Data" needs of clinical and translational research.

Introduction

i2b2[1] (Informatics for Integrating Biology and the Bedside) is an open source software platform widely adopted in academic medical centers across the country. It provides a user friendly front-end interface defining the inclusion criteria for patient cohort identification, which is implemented by the ingestion of clinical data points into the i2b2 relational database, as facts to fulfill those queries on the back end. Typical data dimensions include demographic information and clinical data from administration and electronic medical record (EMR) systems, and billing data from financial systems, as well as tissue and genetic information from research bio-repository systems. As both the data volume and heterogeneity simultaneously grow, finding economical solutions for storing and querying such data becomes an increasingly important concern. Apache Hadoop[2] offers solutions for meeting these challenges by facilitating distributed data processing on relatively inexpensive commodity computing clusters. Using Hive[3] as a supplement for i2b2's SQL engine provides an efficient and economical method of scaling the backend database across multiple computers while completely eliminating the technical burden of implementing customized parallelization techniques like MapReduce.

Method

The Hadoop ecosystem offers HiveQL as a SQL solution, allowing the same SQL statements as in a relational database to be executed against the Hive tables in parallel across the whole cluster at the same time. This feature can be ported directly into the i2b2 CRC cell, so that the queries generated from the generic i2b2 template can be re-directed to the Hive run-time components on the Hadoop cluster. More specifically, we port the i2b2 observation facts, the dimension tables and i2b2 ontology tables onto Hive databases while keeping the user and project management operations intact, as in traditional SQL implementation. This strategy not only maintains the integrity and security features of the original i2b2 design, but also reduces the parallelization overhead on jobs that search modestly sized tables, effectively focusing the parallelization efforts on more computationally intensive queries. The original obfuscation procedure is applied after the results return from Hive jobs to reduce re-identification risk.

Result and Discussion

In this presentation, we summarize an efficient and economical method of scaling the back-end database of i2b2 using Hive on the Hadoop stack in a way that seamlessly integrates the parallel and serial database schemas with i2b2 front end. This method takes the best of the serial and parallel back-ends, and integrates them into a hybrid system that allows all of the front-end and middle-ware features to function as they were originally designed. In the described system the clinical data, which accounts for the vast bulk of the data, has been migrated to a Hadoop stack. Our solution allows the remaining relational data (such as admin data and metadata) to be stored in the original i2b2 datastore (for example Oracle or SQLServer) at much reduced cost. This allows modestly sized organizations to scale up the i2b2 database without the otherwise prohibitive cost of an enterprise scale commercial solution.

Reference

[1] https://www.i2b2.org/
[2] http://hadoop.apache.org/
[3] http://hive.apache.org/

Transforming Research Program Management: From a Ticketing System to a Computerized Research Record (CoRR)

Peter J. Embi, MD, MS; Marcelo Lopetegui, MD; Tara B. Borlawsky, MA; Frank Lamantia; Robert Rice, PhD
Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

Abstract
The rapid evolution of interdisciplinary clinical and translational research teams has required the transition from a standard ticketing system for managing research service requests to a comprehensive electronic research record. We report upon process and adoption improvements resulting from the utilization of user centered design and agile programming methodologies to design and develop the Computerized Research Record (CoRR).

Background
The Ohio State University (OSU) was awarded a Clinical and Translational Science Award (CTSA) in 2008. As a CTSA awardee, the OSU Center for Clinical and Translational Science (CCTS) was expected to provide a diverse offering of research services through 13 distinct programs. Mandated metrics required that the CCTS automate the incoming requests for services and staff response. In response to this requirement, the Biomedical Informatics core developed a simple project-based ticketing system called the "Front Door", which was launched in 2010. However, it soon became apparent that both the research and service provider communities required increased functionality. To meet these demands, the "Front Door" was transformed from a ticket tracking application to a more user-friendly and comprehensive Computerized Research Record (CoRR). Using an electronic health record (EHR) metaphor, the requirements of the researcher ("patient") and service provider staff ("clinical staff") were integrated into this new software application. The first version of CoRR was released to the OSU research community in 2012. Since its release, we have continued to develop CoRR into a generalizable solution for both researchers and providers by seeking input from and delivering features requested by both communities. We report here the results of the eighteen-month experience of expanding and utilizing CoRR, and the methods by which CoRR is becoming an accepted platform in the software toolkit of researchers and providers alike.

Methods
Focus groups were used to develop user requirements both for the "Front Door" and CoRR. Pre- and post-release user acceptance was performed on CoRR. Several methodologies were used to gather input from the two primary constituent groups using CoRR: (1) user utilization and attitudes were solicited by surveys; (2) user acceptance testing was employed to gather input on the design and functionality of the user interface; and (3) monthly meetings with research and provider staff were scheduled to seek input on desired functionality and reactions to proposed and released features. The development team used agile programming methods and tools, including JIRA (Atlassian, Sydney, Australia) to track user stories, bugs and technical tasks, and Balsamiq® (Sacramento, CA) to generate wireframe mockups of interfaces and workflow processes for new features. Users participated in design discussions before and after coding, and a CoRR steering committee sets priorities on the implementation of new features.

Findings
In the first six months of use, the "Front Door" received requests from 62 projects and 57 distinct researchers. In the last six months, 339 researchers associated with 227 projects were processed through CoRR. In the same timeframe, the "Front Door" generated 67 requests for services compared to 449 in CoRR. In the past 12 months among CoRR users (research community), the reported one-time users of CoRR has dropped 14%. During the same time period, the percent of users finding CoRR difficult to navigate dropped from 20% to 1.5%. Less than 1% of users report using a comparable software application, and 66% want more project-based functionality. In the past year, 70% of respondents reported improved navigation and 60% report an enhanced user experience.

Conclusions
We have found that user acceptance can be enhanced through the use of agile application development coupled with user-centric design methods. The employment of these tools leads to a perception of continuing process improvement that facilitates user acceptance and on-going use of the application. Next steps include the inclusion of additional OSU service providers and creation of a stakeholder executive committee.

Acknowledgements
This work was supported in part by a CTSA grant from the NIH/NCATS (UL1TR000090).

Open Proposals: A Pre-Competitive Interactive Space for the Research Community

Oksana Gologorskaya, MS, Mini Kahlon, PhD, Leslie Yuan, MPH, Rachael Sak, RN, MPH, Cynthia Piontkowski, Lisa Schoonerman, Courtney McFall, Clinical & Translational Science Institute, University of California, San Francisco, San Francisco, CA, USA

Summary

UCSF Open Proposals (OP) is an online platform for open and collaborative proposal development for research funding and innovation initiatives. The process, enabled via a web-based system designed by CTSI at UCSF, provides an interactive pre-competitive space for researchers and administrators to share and discuss ideas and proposals before submitting them for review. Compared to traditional black box submissions/review, this model encourages pre-review input and commenting from the wider interdisciplinary community, enables creation of stronger teams, and produces higher quality proposals due to the ability to revise proposals based on feedback prior to final submission. In the last year and a half, the Open Proposals process has been used 15 times, helping disburse a total of about $3 million.

Introduction and Background

Standard proposal submission processes do not enable the development of the best projects and teams.

Biomedicine, research, and research administration projects are often funded through processes that involve requesting proposals from a community and evaluating their merits through a peer-review process. However, black box reviews can sometimes include redundant proposals and offer insufficient opportunities for collaboration and external input.

Open Proposals enables an interactive online process for open, collaborative proposal and team development.

UCSF implemented OP over 4 years ago and to date, 19 OP opportunities have been run, covering a variety of topics, including the solicitation, improvement and selection of ideas for -

- *New initiatives for the CTSI's $112 million renewal proposal to NIH*
- *New hi-tech cores at UCSF from a $2 million pool*
- *Projects to improve research administration at UCSF from a $350K pool*
- *Ideas for the 'Caring Wisely' initiative to improve value of healthcare delivery.*

Methods

The Open Proposals process requires submitters to post their proposals online where they are available for review and comment by a community of peers. Each OP opportunity includes three phases. The first phase is "Submission", where proposals are posted and open to the broadest possible community. Submitters may post preliminary or fully fleshed-out ideas for comment and/or to seek collaborators to help develop an idea; all proposals must be submitted by an established deadline. The second phase is "Open Improvement", when the community comments upon proposals and/or joins teams. This is what sets the OP approach apart from traditional processes. Submitters can then use the feedback to revise/improve upon their proposal throughout this phase. The third phase is "Review", when public comments and proposal revisions are closed, and proposals are reviewed internally by committee, selected for award, and announced online.

The Open Proposals platform is built on the Drupal CMS workflow engine, and hosts OP opportunities. The platform generates a standalone subsite for every opportunity, featuring user-friendly workflows, and customizable branding for the sponsor organization. The platform is integrated with the University research networking tool, UCSF Profiles.

Standardized surveys of proposers and commenters are conducted and are analyzed to help inform improvements to the process and tool.

Results

Open Proposals improves proposals and teams.

Open Proposals has been successfully adopted at UCSF. Key metrics and indicators we track include:

- *Adoption*: Over the last 1.5 years, 15 initiatives from six UCSF campus groups were run through the Open Proposals process, disbursing a total of about $3 million in funding
- *Participation*: On average, each of the 15 initiatives received 27 proposals and 127 comments (with some proposals receiving as many as 50 comments)
- *Outcomes*: We saw a variety of useful behaviors, including having proposals retracted after the public comment period (showing the ability of the process to organically weed out weaker ideas), and the merging of proposals by different teams (highlighting the ability of the process to expose redundancy).
- *Interest from other institutions:* Several other institutions, including Harvard, the University of Minnesota, and the University of California, Merced have expressed interest in implementing and using Open Proposals. We launched a pilot project with UC Merced to host their strategic opportunity on the UCSF Open Proposals environment.

Analyzing Problem List Data for Real-time Re-Use

Courtney Hebert[1], MD; Chaitanya Shivade[2]; Philip Payne[1], PhD; Peter Embi[1], MD, MS, Department of Biomedical informatics[1]; Department of Computer Science and Engineering[2], The Ohio State University, Columbus, OH

Abstract:
Electronic problem lists (EPL) are a source of real-time comorbidity data available for research and risk-modeling. However, best practices for EPL use in secondary applications are unknown. This study compares the EPL to administrative data and studies how the EPL changes over the hospitalization in order to define best practices.

Introduction:
Administrative billing data have historically been used to detect comorbidities, with moderate sensitivity compared to manual chart review[1], however up-to-date administrative data is often not available until after discharge. Natural language processing techniques have been used to determine comorbidity status but this can be difficult to use in a real-time environment. EHR-based problem lists are a potential source of real-time clinical data, however earlier studies have shown low sensitivity for detecting comorbidities when compared to administrative data.[2]

Background:
In order to operationalize an EHR-based readmission risk tool we required comorbidity data that was available during a hospitalization. We hypothesized that EPL data in our comprehensive EHR would be sufficiently complete and accurate to use for real-time risk prediction. To test this hypothesis we studied the problem lists of a retrospective cohort of congestive heart failure (CHF) patients during a single hospitalization. In order to understand the nature of changes to the EPL, we measured changes in the number of problems present per day in the EPL. We also measured agreement between EPL and administrative data on comorbidity classification to assess its completeness.

Methods:
We collected EPL data as well as administrative discharge billing data for a retrospective cohort of patients discharged from the hospital after an initial admission for CHF from 10/15/11 to 5/1/12. The data were used to establish the number and types of problems on the EPL for each day of hospitalization. To determine whether the content of the EPL approximated administrative data, we classified the comorbidities based on a well-known classification system[3] and calculated Kappa statistics for each major comorbidity.

Results:
565 index admissions for CHF were included in the analysis. 510 (90%) patients had at least one EPL element present during their index admission. Of those with at least one EPL element recorded, the number of problems on the first day of hospitalization ranged from 0-32 (median 8). On the last day of hospitalization the range was 1-33 (median 9). A slight majority of patients (51%) had the same number of problems on their list on the day of admission as on the day of discharge however, 46% of patients had an increase in the number of problems while only 4% had a decrease. Calculated Charlson comorbidity scores were stable over the hospitalization for 75% of patients and increased for 24% of patients. Kappa statistics were in the range of good to excellent reproducibility for select comorbidities including diabetes (0.76), renal failure (0.62), obesity (0.56) and peripheral vascular disease (0.56), however reproducibility varied significantly between comorbidities.

Discussion:
At our institution, there is significant variability in the number of problems on each patient's EPL. During a hospitalization, however, the EPL varies minimally for most people, and the majority of the change involves the addition of new problems. In addition we found that for most major comorbidities there is reasonable agreement between EPL and administrative billing data. This study adds to a limited literature on this subject and is a vital first step in understanding the best way to operationalize real-time risk prediction tools to enable personalized and predictive health care. We are currently exploring whether temporal patterns of EPL changes can aid prediction of health outcomes including readmissions.

References

1. Quan H, Li B, Saunders LD, Parsons GA, Nilsson CI, Alibhai A, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health services research. 2008;43(4):1424-41.
2. Szeto HC, Coleman RK, Gholami P, Hoffman BB, Goldstein MK. Accuracy of computerized outpatient diagnoses in a Veterans Affairs general medicine clinic. The American journal of managed care. 2002;8(1):37-43.
3. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care. 2005;43(11):1130-9.

Design & Implementation of a Real-Time Location Sensing System for Determining Interpersonal Social Contacts in the Emergency Department

Sarah A Hilton, MSHS[1], Douglas W Lowery-North, MD, MSPH[2], Vicki S Hertzberg, PhD[2], Lisa K Elon, MS, MPH[2], George A Cotsonis, MS[2]
[1]Medical College of Georgia at Georgia Regents University, Augusta, GA 30912
[2]Emory University, Atlanta, GA 30322

Summary: We describe the planning, acquisition, and implementation of a radiofrequency identification (RFID) system for tracking close proximity interactions (CPI) in an emergency department (ED), and its integration with electronic health records (EHR) to model potential cross infection among patients and staff.

Introduction: EDs facilitate the comingling of patients with acutely infectious conditions and patients highly susceptible to. The presentation of an infectious patient to the ED represents a major risk for cross infection. Better understanding of the potential routes of transmission could lead to improved countermeasures for cross infection in healthcare settings.

Background: Although EHRs can provide abundant amounts of clinical data they fall short of identifying the geographic and social contexts that could lead to cross infection. Linking data from an RFID system and an EHR provides a tool to assess the dangers associated with the management of seriously infectious patients in EDs. Our objective was to quantify the interactions between and within different segments of an ED's population, using influenza as an exemplary pathogen for cross infection.

Methods: We designed hardware, software and data models to facilitate the integration of individual patient information with real-time location data during emergency care in the ED of Emory University Hospital Midtown, Atlanta, Georgia, an urban, academic hospital with an annual ED census of 57000 patient visits. We scaled the RFID system to detect CPIs within the one meter radius characteristic of airborne influenza transmission. We abstracted data from the EHR and integrated them with RFID data to determine CPIs.

Results & Discussion: We deployed the RFID system from 1 July 2009 to 30 June 2010. We intended to measure 104 12-hour shifts during this period. 23 study periods had to be removed from analysis due to equipment issues, staffing, and inclement weather. One third of study periods had to be shortened due to similar issues. Other studies have experienced similar difficulties to the same extent. Tag location events were measured in 94 zones in the 25,000 square foot ED. 100% of patient location events were linked to the EHR. We measured 277,858 tag location events during 17,001 tagged patient hours from 4,731patient visits. Location information was used to discern 293,181CPIs. This novel data model facilitates the use of social network analysis (SNA), as illustrated in Figure 1. SNA measures such as degree, closeness, and betweenness can be used in regression analysis to examine differences between staff categories as well as patient complaint groupings. Identification of locations where many CPIs occur and/or staff categories involved in many CPIs can be used to support development of countermeasures to cross infection.

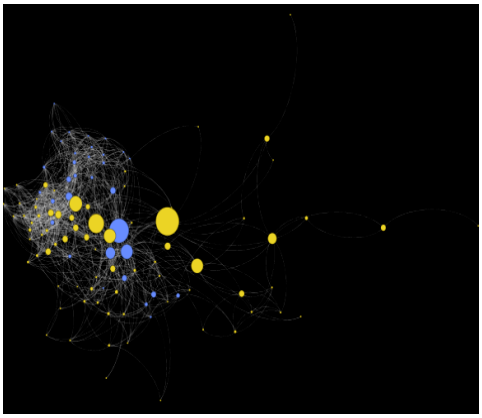


Figure 1: Social Network Graph of CPIs Among Patients (Yellow) and Staff (Blue)

Natural Language Processing of Free-text Problem List Sections in Structured Clinical Documents: a Case Study at NIH Clinical Center

Vojtech Huser, MD[1], PhD, Kin Wah Fung, MD, MS, MA[2], James J. Cimino, MD [1,2]
[1]Laboratory for Informatics Development, NIH Clinical Center, Bethesda, MD
[2]National Library of Medicine, Bethesda, MD

Abstract

Maintaining an electronic problem list is important for managing patient care and also important in numerous research analyses within healthcare integrated data repositories (IDRs). We used MetaMap natural language processing (NLP) tool to detect SNOMED controlled terms in free-text clinical documents with a structured problem list section. We present findings on copy rate of problem list sections, detected concepts terms and NLP tool performance, relevance of NLM CORE problem list set and comparison with ICD9CM coder's data.

Institutional background: We used integrated data repository of the intramural research program at the National Institutes of Health (NIH) called BTRIS (Biomedical Translational Research Information System). This repository contains data from the electronic health record (EHR) system of the NIH Clinical Center, a 240-bed hospital devoted exclusively to research, as well as other clinical and research systems.

Methods: We queried the BTRIS database to establish a cohort of deceased patients with free-text problem list data. Deceased patient data was used to conduct a pilot experiment to arrive at optimal NLP tool configuration that could be later used for converting free-text problem lists of all CC patients to coded data. Extracted problem list (PL) section data was optimized for input into MetaMap NLP tool by eliminating duplicate identical sections. We also wanted to estimate the copy-forward rate (frequency of clinician's cutting and pasting of previous PL section text). A significantly high copy-forward rate can be an incentive for clinicians to adopt electronic problem list. To consider physician adoption of detected coded problems, we want to determine an optimal targeted subset of coded problem list terms. We evaluated how often the detected SNOMED codes are present in NLM's CORE Problem List Subset of SNOMED CT (version 2013-02). In addition to arriving at optimal NLP configuration, we compared by limited manual review codes detected by NLP in clinician's problem list sections with ICD-9-CM codes generated by medical coders.

Preliminary Results: Within the 24 355 deceased patients in our repository, a total of 1062 had at least two clinical documents with a problem list section. Documents authored within CC's EHR enable identification of identical clinical document sections (e.g., problem list section) across different document types. The top 3 most frequent document types with a populated problem list section were 'Standard SOAP Progress Note', 'Discharge: Routine Progress Note' and 'Cardiology SOAP Progress Note'. Due to significant NLP computing time required to parse all PL sections, we eliminated repeating identical PL sections within each patient. The average number of PL sections per patient was 27.01 (median 7) when repeating sections were counted. Considering only unique section, the average reduced to 9.73 PL section (median 4) per patient. The mean copy rate (percentage of copied PL sections) per patient was 42%. We used MetaMap NLP tool to parse all PL sections with settings to only detect SNOMED CT terms. The median number of MetaMap detected concepts was 11 per patient (range 1-388). MetaMap detected a total of 2315 unique SNOMED concepts of which 50.9% (1178 concepts) were in the CORE subset. In the coders' ICD data, we found on average 10.24 distinct ICD codes per patient (median: 5). Data on detected SNOMED concepts (within and outside CORE), PL sections examples and other on-line data appendices are available at http://dx.doi.org/10.6084/m9.figshare.808592. Limited qualitative comparison of diagnoses from coders versus clinicians (after NLP) (on individual patient level) indicates more granular diagnoses in the clinician's PL sections, but greater breath of coverage and consistency in coders' data.

Conclusion: Although prior studies in assigning diagnostic codes exist [1], our study is the first to focus exclusively on problem list sections of semi-structured clinical documents. Our findings are relevant to organizations interested in pre-populating an EHR problem list module (mandated by meaningful use criteria) with data from NLP parsing of retrospective semi-structured clinical documents or improving their existing PL modules (optimal PL valueset).

References

1. Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc 2006;13:516-25.

LabKey Server: An Open Source Platform for Large-Scale, Translational Research

Mark Igra, BS[1], Elizabeth K. Nelson, PhD[1], Britt Piehler, MBA[1], Josh Eckels, BS[1], Matthew Bellew, BS[1], Peter Hussey, BS[1], Adam Rauch, BS[1]
[1]LabKey Software, Seattle, WA

Abstract

Translational research teams must make sense of diverse types of data while collaborating across boundaries of systems, organizations, and expertise. The LabKey Server open source platform (http://labkey.org) helps such teams integrate, reproducibly analyze, and securely share not just clinical and specimen information, but also the complex results of high-throughput assays.

Background

Organizations engaged in large-scale, translational research face a daunting array of data management challenges. To speed progress towards disease therapies, researchers need software systems that help them manage not just clinical and specimen data, but also the flood of complex information produced by modern molecular and cellular techniques. Frequently, data must be gathered from distributed sites in a standardized manner, screened for quality, linked with other data sources, and securely shared across research teams. Collaborating scientists with varied expertise must be able to access, analyze and visualize pooled data, both to interpret trials under way and to generate new hypotheses for further studies. Upon publication, organizations need to offer public access to the data, metadata, and analytical tools such that independent reviewers can reproduce, validate and extend research results.

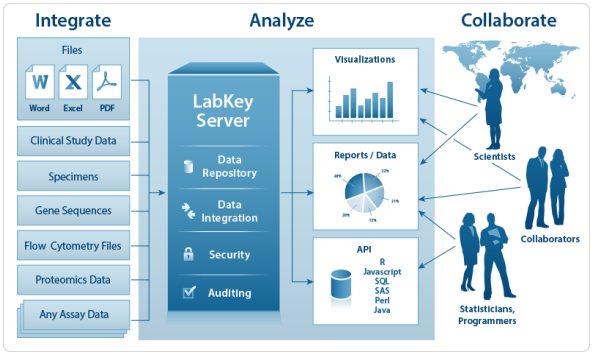
We developed LabKey Server to address these challenging requirements of translational research.

Methods/Results

LabKey Server is a Java-based web application that runs on the Apache Tomcat web server and stores its data in a relational database.

It includes tools for electronic data capture, a file management system that provides search capabilities, and a flexible, relational data store that supports integration and aggregation of results. A secure query service ensures that users can only view data that they have permission to see, no matter how they access the data (via the API, visualizations, SQL queries, or otherwise). Users can define new data types via the user interface, as well as associated metadata and wizards for data collection. Visualization and analysis tools include a graphical view designer and a built-in R interface. API libraries in a variety of languages support customization and extension. Developers can add modules or simply include API-enhanced content (*e.g.*, R visualizations) in wikis or HTML pages. LabKey Server readily integrates with and extends systems that are based on databases and/or spreadsheets.

Discussion



LabKey Server acts as a web-based portal for integration, analysis and secure sharing of the diverse data types produced by translational research. The system supports annotation and acquisition of assay results in a standardized manner, helping teams scale up from pilots to trials. Analyses and visualizations can be shared and collaboratively evolved via the web portal. Data de-identification tools support public data sharing post-publication. Primary study data can be reused and refreshed in ancillary studies.

Over 100 installations of the LabKey Server platform serve translational research organizations all over the world, including the Statistical Center for HIV & AIDS Research at the Fred Hutchinson Cancer Research Center (Atlas: http://atlas.scharp.org) and the Immune Tolerance Network (ITN TrialShare: http://itntrialshare.org). In August 2013, ITN used ITN TrialShare to break new ground for clinical trial transparency. ITN directly linked a major clinical trial publication to participant-level data and interactive analyses on the TrialShare web portal, enabling immediate, public exploration of results.

Source code, compiled binaries, and documentation, are professionally maintained and freely available under the Apache 2.0 license at http://labkey.org.

SNAPL-CAT: Catalyzing the rate-limiting step of big data psychometrics with item-response theory and advanced computerized adaptive testing

Ming-Chih Kao, PhD, MD, Stanford Hospital and Clinics, Karon Cook, PhD, Northwestern University, Garrick Olson, BS, Teresa Pacht, BS, Beth Darnall, PhD,Susan C. Weber, PhD, Sean Mackey, MD, PhD, Stanford University, Stanford Hospital and Clinics

No disclosures, Funding NIH HHSN 271201200728P and Redlich Pain Endowment

Introduction: Unlike passive biometric measurements, psychometric measures require active participation from subjects and are rate-limited by subject burden. NIH PROMIS and Toolbox provide efficient questionnaires based on item-response theory (IRT) and computerized adaptive testing (CAT). Leveraging these item banks and developing advanced CAT algorithms can reduce burden and enable big data psychometrics.

Methods: The algorithm SNAPL-CAT is developed on open source MEAN stack (MongoDB, Express, AngularJS, and NodeJS) with D3.js visualization. Item banks and linkages are obtained from Northwestern Access Center and PROsetta Stone. Features include initialization (individualized or patient population priors), item selection (expected Kullback-Leibler, minimum expected posterior variance), advanced item selection (alpha-stratification, exposure control, content balancing, probabilistic constrained optimization), stopping rule (predicted standard error reduction, percentile width, hybrid), and estimation (expected a posteriori, maximum likelihood, maximum a posterior). Performance in 4,466 measurements in the Stanford-NIH Pain Registry are analyzed.

Results: Basic CAT provided significant reduction in burden (mean number of items ± SD, fold reduction): Anger (6.24+/-1.21, 4.6-fold vs BPAQ), Anxiety (4.93+/-0.97, 1.4-fold vs GAD-7), Depression (4.97+/-1.07, 1.8-fold vs PHQ-9), Fatigue (4.78+/-0.76, 8.4-fold vs FACIT-F), Physical Function (4.11+/-0.48, 4.9-fold vs HAQ-DI), Pain Interference (4.19+/-0.71, 1.7-fold vs BPI), Sleep Disturbance (4.95+/-1.41, 2.4-fold vs SDQ), Sleep-Related Impairment (4.54+/-1.24, 1.8-fold vs ESS). Altogether, tfhe 132 classic instrument items may be alternatively assessed by 38.7 +/- 7.9 items, for 2.8 to 4.3 fold reduction in patient burden.

Conclusions: Using advanced IRT and CAT, the Stanford-NIH Pain Registry and SNAPL-CAT leverage the powers of NIH PROMIS and Toolbox, and enable big data psychometrics.

Importing Continuity of Care Documents into i2b2 and SMART

Jeffrey G. Klann, PhD [1,2], Alyssa Porter, MS [1], Nich Wattanasin, MS [1], Shawn N. Murphy, MD, PhD [1,2]
[1] Partners Healthcare System, Boston, MA, USA
[2] Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA, USA

The Meaningful Use incentive program will require the Consolidated Clinical Document Architecture (C-CDA), an HL7 standard for electronic clinical data. Therefore, C-CDA-formatted patient data will soon become widely available. Here, we describe an Integrating Biology and the Bedside (i2b2) module to import these documents, for populating the data repository and refreshing patient data to reflect near-real-time information. This can be utilized by SMART (Substitutable Medical Apps, Reusable Technologies) apps, such as SMART-i2b2's clinical trial platform.

Background

The Consolidated Clinical Document Architecture (C-CDA) is an HL7 standard for expressing patient data electronically, and the Office of the National Coordinator for Health Information Technology (ONC) selected it as a requirement for Stage 2 of the Meaningful Use incentive program. Therefore, healthcare facilities will shortly be producing large numbers of these documents with up-to-the-minute patient information. Partners Healthcare in Boston, MA is building the capacity to generate at least 75,000 documents per day.

Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH National Centers for Biomedical Computing. It is a flexible, componentized clinical research and data warehousing system that now enjoys widespread adoption at over 80 sites nationwide.

Substitutable Medical Apps, Reusable Technologies (SMART) is an ONC-funded platform for reusable medical apps that can run on participating platforms connected to their various electronic health records. SMART is fully integrated in i2b2, supporting "deep dives" into the patient record directly from i2b2. [1]

Adding C-CDA import to i2b2 will create a standards-based import process from a readily available data source, and it can be used for a "live refresh" of individual patient data in SMART.

Methods

To support C-CDA import into i2b2, we developed the SETL (Service-Based Extract, Transform, and Load) cell to convert C-CDA documents into i2b2 format. We also modified the SMART cell to optionally retrieve up-to-the-minute data from the SETL cell rather than from the data repository. Finally, we created an i2b2 ontology for C-CDA documents so that the converted document can be stored directly in the data repository.

Because C-CDA is so expressive, it is possible to have very different C-CDAs that contain the same information. Therefore, we turned to the Open Health Tools (OHT) organization to help us create a solution that would require minimal modification across organizations. OHT is responsible for the official C-CDA validator used in Meaningful Use. OHT is working on C-CDA translation tools that are resilient to organizational differences in expression.

For this first release of the SETL cell, we targeted the sections of the C-CDA Continuity of Care document profile (CCD) needed for our clinical trial recruitment SMART application: demographics, problems, and notes.

The SETL cell is open source and will be available on the i2b2 wiki in April 2014.

Results

For this release we elected to implement and test our SETL cell at Partners Healthcare, which provides a robust C-CDA generation service from outpatient encounters across the enterprise. We have developed an intermediary program that, at the SETL cell's request, calls upon the service to retrieve a document for a given patient. This intermediary program also adds clinical notes into the document (Meaningful Use does not require clinical notes).

The SETL cell is running at Partners Healthcare and will be part of SMART-i2b2's clinical trial platform.

Discussion

We have achieved import/refresh of up-to-the-minute patient information in i2b2 and SMART by leveraging C-CDA. Although we have only tested the SETL cell against Partners Healthcare documents, we expect our design will make adaptation to other sites straightforward. Future work includes supporting all sections of Meaningful Use C-CDA documents. C-CDA will likely become an important data source for secondary use of clinical data.

Thanks to the MDMI team (part of OHT), Martin Rees, and Mike Buck (at the New York City Department of Health and Mental Hygiene). Sponsored by ONC 90TR001/01.

References

1 Wattanasin N, Porter A, Ubaha S, *et al.* Apps to display patient data, making SMART available in the i2b2 platform. In: *Proceedings of the AMIA Symposium*. 2012.

A Scalable Approach to Dynamically Populating REDCap-enabled Research Registries from an Enterprise Data Warehouse

Frank J. Lamantia; Robert R. Rice, PhD; David Ervin; William Stephens; Gregory Young, MS; Tara B. Borlawsky, MA

The Ohio State University, Department of Biomedical Informatics, Columbus, OH

Abstract

To address the increased demand for custom research registries in a timely and streamlined manner, we developed a scalable, modular architecture for integrating clinical data from an enterprise data warehouse with an off-the-shelf electronic data collection tool (REDCap). We describe our development objectives and system implementation.

Introduction

The demand for research registries comprised of data collected during standard clinical care has increased. The OSU Information Warehouse (IW) is an enterprise data warehouse that curates data from multiple information systems and contains datasets such as patient management, billing and finance, procedures, medications, laboratory results, reports, order entry, outcomes and demographics[1]. The three underlying goals for the implementation of this architecture are to: (1) introduce a standardized workflow (human and technical) to meet the majority of the research community needs in an expedited-time frame; (2) create a readable, fully discrete datamart implementation that can be leveraged across initiatives, and (3) provide the research end users with a graphical user interface for defining registry-specific customizations, as well as the use of a familiar application such as REDCap[2] to visualize the data.

Technical Architecture

The infrastructure that we have developed to support research registries is comprised of the following core components (Figure 1):

(1) A common registry datamart, based upon the structure and content of the Observational Medical Outcomes Partnership (OMOP) common data model[3], has been developed within the IW to expose and aggregate a core set of frequently requested data elements, including demographics, diagnoses, procedures, medications and laboratory results.

(2) A lay-friendly user interface has been developed that allows the researcher to customize their registry by selecting which subset of the common data elements they would like to be included. In addition, the researcher can define the criteria for identifying and including follow up encounters.

(3) The REDCap data dictionary is dynamically generated based upon metadata from the IW for the selected data elements (Step 2), and facilitates the transfer of data into a REDCap project associated with the registry.

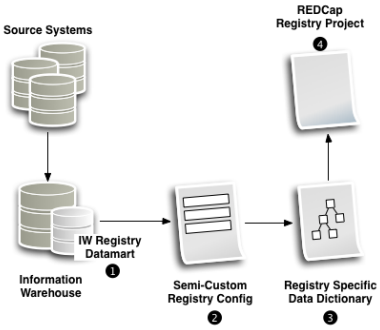


Figure 1. Overview of architecture supporting the integration of EHR data with research registries.

(4) The final component of the architecture is a REDCap project in which the research team can document informed consent and collect research-specific data, as well as view and abstract EHR-data extracted from the IW.

Discussion

From a technical standpoint, the datamart component (1) shields the complexity of a highly-normalized table structure and often ambiguous naming standards from the researcher. From a logistics standpoint, project-specific datamarts can take 4-12 weeks to implement. By defining a standard structure comprised of a core set of data elements and generic data transfer pipeline, we are able to streamline the process of standing up a new registry.

Conclusion

While the current architecture supports the integration of basic clinical data with research registries, future work aims at addressing the following: discrete collection of domain-specific clinical data via the EHR and subsequent integration with the REDCap-based registry, and tighter integration with statistics packages.

References

1. Kamal J, et. al. Information warehouse - a comprehensive informatics platform for business, clinical, and research applications. AMIA Annu Symp Proc. 2010 Nov 13;2010:452-6.
2. Harris PA, et. al. Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377-81.
3. Overhage JM, Ryan PB, Reich CG et al. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60.

Joint Analysis of Multiple Data Types in Electronic Health Records

Joseph E. Lucas, Ph. D.[1,2]
[1]Quintiles, Morrisville, NC; [2]Duke University, Durham, NC, USA

One of the great challenges in creating applications based on electronic health records (EHRs) is the diversity of data. Free text, continuous values and binary data are among the many different types of data available in EHRs. There is a large body of work built around utilizing particular data sources within the record to address specific medical applications. However, to date there has been little effort spent in their joint analysis. Modeling a single type of data presents problems for the general applicability of statistical and machine learning approaches, particularly in cases when critical information comes from multiple EHR sources.

We have developed a Bayesian statistical model for patient clustering that enables simultaneous modeling of essentially arbitrary collections of text, continuous and binary variables. In addition, we develop an inference approach for fitting this model that does not require the sampling of any nuisance parameters. This facilitates use of the model on large data sets.

Methods. Our data consists of approximately 50,000 emergency department visits. From these visits, we use 5 separate EHR documents: patient reported medications, written diagnosis (plain text), chief complaint, orders and primary nurses' notes. We incorporate 6 continuous variables: age, systolic and diastolic blood pressure, temperature, pulse and respiratory rate. Finally, we include gender and disposition (treated and released versus admitted to hospital) as binary variables.

		Model			
Experiment	Test	Chief Complaint	NLP concepts	Cluster	NLP + Cluster
Drug versus patient subgroup	% that validate	18%	22%	58%	56%
Differential diagnosis	Average correlation, train versus test	0.26	NA	0.62	0.61

Table 1: Our model based clustering approach - with either raw or NLP processed data - leads to higher rates of validation in both experiments. 58% and 56% of drug-subpopulation relationships validate versus 18% and 22% for comparison models. Correlation of differential diagnosis probabilities between training and test data sets (averaged across all subgroups) is also much higher for our model regardless of preprocessing.

We split our data into training and test data sets in order to assess the ability of our model to consistently generate homogeneous patient clusters. In all cases, the clustering model was learned on the training data and directly applied to the test data. Homogeneity was tested in two ways: (1) We tested for statistical relationships between drugs and patient clusters in the training data, without regard to the reason for the association (Fisher's exact test of drugs by subgroup), and computed the percentage of these relationships that validated, (2) we computed the differential diagnosis of patients in each subgroup in the training set and compared to the test set (Kendall correlation of diagnosis probabilities). Performance in each test was compared for four different approaches to patient clustering: (1) our model as described, (2) clustering by chief complaint, (3) patient subgroups based on MetaMap NLP (natural language processing) concepts, (4) our model after exchanging the "raw" nurses' notes for NLP processed concepts learned from those notes (a clustering-NLP hybrid model).

Results. Our model based clustering approach with either NLP processed or raw nurses' notes produced results that validated 2-3 times more often than either of the other approaches in the "Drug versus Subpopulation" experiment (see Table 1). They similarly outperformed chief complaint on the "Differential Diagnosis" experiment. Assignment of differential diagnosis based on only NLP concepts is not straightforward, therefore this test was not done. These results demonstrate the power of incorporating multiple sources of information about patient status.

Discussion. With the statistical model outlined in this work, we have demonstrated an ability to pull together widely varying and critically important sources of information in EHRs. Identifying groups of patients with similar disease states, as our model does, makes possible an array of important medical applications. Any new patient belonging to such a group will share many clinically relevant group characteristics including etiology, diagnosis, prognosis, treatment options, and future costs.

Workflows for a Web-based Consent Management System with Electronic Consent

Keith Marsolo, PhD and Jeremy Nix
Cincinnati Children's Hospital Medical Center, Cincinnati, OH

Abstract
In an attempt to decrease the transaction costs for patients wishing to participate in research, we have developed a web-based consent management system that includes both e-consent and consent tracking modules. It supports multi-center studies, with protections to ensure that centers have access to patient-level data only for patients at their center.

Introduction
The idea of using web-based or electronic consent (e-consent) to lower the transaction costs for patients wishing to participate in research without sacrificing the process of informed consent is one that is gaining traction within the informatics community. We present a system that was originally developed to support minimal risk, observational pediatric research studies that also include a quality improvement (QI) component. In these studies, the full patient population is often included for QI, while a smaller subpopulation has consented for research. Its use in pediatrics requires the implementation of workflows for parental consent, parental consent and patient assent and patient consent.

Methods
Participants can be added and maintained within the system either individually or through a bulk upload process. The application stores demographic information on participants. It tracks a patient's consent status and can fire alerts to the study staff when the patient's consent expires (when they turn 18, for instance). Once a patient's record has been created in the application, patients can be registered in either a paper or e-consent workflow.

The e-consent process begins with potential participants receiving a recruitment that includes a link to the e-consent web site, along with an invitation code. The potential participant accesses the e-consent website and provides a secondary piece of data for identity verification purposes. They are then able to review the consent document via the website and can even leave the website to discuss the research with others, all prior to actually providing consent. Once completed, a copy of the fully executed form is e-mailed to the patient as well as to the study staff. If the patient is under the age of majority, a modified workflow is followed that can allow for both documented assent and parental permission. The e-consent module also includes functionality for participants to e-mail study staff with questions. This functionality is not only convenient for the participants but also provides the study staff with a tool for maintaining a record of their interactions with the participant during the initial consenting process.

Since the consent management system contains a mapping between study-specific identifiers and patient identifiers, another major use case of the tool is to act as a service that can be queried to replace study identifiers with patient names or medical record numbers (or vice versa) to produce reports that are used to support clinical care processes, such as pre-visit planning and population management reports.

Results
Through a soft launch, the consent management system has been trialed by seven studies at Cincinnati Children's. Four studies are using the e-consent process as a way to capture the patient's signature electronically, having patients go through the process during the clinic visit. Two have consented patients via e-mail, though they report having to spend a fair amount of time chasing down patient responses. One study is simply using the service to replace study ids with patient identifiers on their reports.

Discussion
The consent management system has provided a way to lower the administrative overhead imposed upon clinical research coordinators by the process of informed consent. It has allowed patients at satellite clinics to participate in research projects for which they might not otherwise be recruited. We are exploring ways to increase patient response to the recruitment e-mails and plan to study patient's perception of the e-consenting process as well as the on their understanding of the study.

A Data Set Authoring Tool and Code Generator for Secondary Analysis in Distributed Research Networks

Daniella Meeker, PhD[1], Christopher Skeels [1], Laura Pearlman[2], Karl Czajkowski[2], Lucila Ohno-Machado, MD, PhD[3]

[1]RAND Corporation, Santa Monica, CA, [2]University of Southern California Information Sciences Institute, Marina Del Rey, CA, and [3]Division of Biomedical Informatics, University of California San Diego, La Jolla, CA

Abstract

We developed a tool to facilitate comparative effectiveness research in clinical distributed research networks. The tool, developed as part of SCANNER (Scalable National Network for Effectiveness Research), allows researchers to define rules for processing data from clinical sources, and generates SQL code for preparing analysis data sets from transactional data.

Motivation

Data processing may have more impact in inferences made from secondary analyses than particular analytic or estimation methods. Investigators participating in multisite research are often faced with the tradeoff between depending on staff at partnering sites to prepare data, or facing delays associated with legal agreements and complex IRB reviews. To support investigators in the SCAlable National Network for Comparative Effectiveness (SCANNER), three related needs were identified: (1) investigators unfamiliar with database programming required a graphical user interface for developing well-defined specifications for data processing rules for new studies; (2) sites that had invested in data standardization should be provided with executable programs for data processing; (3) rules should be reusable and discoverable across different studies. We used two studies in the SCANNER portfolio as test cases for determining how to meet these needs. While several excellent tools and software for building reports and distributed queries were available, our test cases required support for more complex data processing rules and more flexible SQL generation capabilities than they could provide.

Materials and Methods

We identified the following gaps in existing tools that could meaningfully support the process of "bringing the questions to the data": (1) tools with the most sophisticated capabilities for specifying data processing rules did not generate specifications that could be translated into computable queries; (2) tools for authoring distributed queries did not support processing with sufficient complexity to produce analytic data sets for our test cases; (3) report generation software required direct linkage to underlying data models and did not generate human readable specifications that could be distributed independently or feasibly translated to other sources.

Results and Conclusion

We developed a system that incorporated several features from existing tools to develop a graphical data set builder for investigators. We employed the Value Set Authority Center and Observational Medical Outcomes Partnership (OMOP) V4 Vocabulary tables to incorporate most National Library of Medicine standards for terminology translation and code groupings to simplify the process of identifying code sets. The National Quality Forum's Quality Data Model rule set was implemented into the user interface for specifying derived variables related to temporal patterns and logical relationships in clinical data. A set of rules for specifying conversion from transactional data to de-identified patient-level data sets was developed and incorporated into the user interface.

Specifications for data processing were represented in an emerging standard used for EHR meaningful use certification derived from the Health Quality Measure format. A translation engine for converting these data to SQL was implemented, with a plug-in for transforming these specifications into queries against the OMOP V4 data model. The tool is linked via web services to the SCANNER data set registry so that variables, programs, and data sets can be reused in related studies and incorporated into study protocols for distribution to participating clinical sites.

We demonstrated the feasibility of automating some portions of data pre-processing for comparative effectiveness research in two use cases. We are currently evaluating the application of this tool in other SCANNER use cases.

Acknowledgements: We thank AHRQ support from R01HS019913 and the members of the SCANNER team.

Discovery of Seasonal Patterns in Incidence of Disease from EHR Data

Rachel D. Melamed, MPhil, Hossein Khiabanian, PhD, and Raul Rabadan, PhD
Department of Biomedical Informatics and Department of Systems Biology
Columbia University College of Physicians and Surgeons, New York, NY, United States

Summary: We leverage patient diagnosis information from an Electronic Health Records (EHR) database to identify diseases with periodic patterns in incidence over time. By accounting for observed biases in this data, our method is able to isolate a small set of diseases with seasonally increased risk, including a novel seasonal pattern in acute exacerbation of myasthenia gravis.

Introduction and Background: The wide diversity of human disease, and many details of disease cases, is increasingly being compiled in EHR. As these data grow in size and quality, methods that can extract meaningful patterns from them will have the power to uncover new insights into the causes or consequences of illness. One illuminating aspect of human disease is the distribution of incidence over time; finding that cases of a disease cluster in certain time intervals can potentially inform us of unconsidered causes of these cases. No method has systematically examined incidence of diagnosis to find evidence of periodic patterns and distinguish these from confounding temporal signal.

Methods: We utilize the New York Presbyterian EHR, a system with hundreds of millions of records, extracting 2,800 ICD-9 codes with greater than 500 diagnoses from 1997 to 2009. Compiling the number of diagnoses per month, two confounding characteristics of the temporal data appear. These are evident when the hospital population is considered in total. First, there is an increase in number of diagnoses over time, reflecting many effects including changes in the population. To correct for this large-scale trend, we calculate the smoothed incidence of each diagnosis using kernel density estimation, based on the observed diagnosis rates, and we subtract the kernel-estimated density at every month from the actual observed incidence. When the data is corrected for this long term trend, a secondary bias appears: there is a seasonal variation in the aggregate number of hospital visits, and the more frequently a diagnosis is made, the more it correlates with this seasonal variation. It seems likely that a variety of factors influence the seasonal increases in hospital visits in the spring and fall, and these (mostly chronic) diseases are in high enough proportion of the general population that they appear to follow a seasonal trend. In order to correct for this influence, we estimate the number of monthly diagnoses that would occur if a disease was always a fixed proportion of the total diagnoses, by mean-scaling the summed total diagnoses. After subtracting the contribution of the seasonal trend, we have an estimated corrected number of diagnoses per month. We quantify periodicity of incidence using Lomb-Scargle periodograms, a least squares method that assesses predictive power of a range of periods and provides an associated approximate significance.

Results and discussion: Of 2,800 diagnoses examined, fewer than 10% have significant periodic signal, with a variety of patterns as shown in Figure 1A, a variety of patterns appear. Unsurprisingly, we find that viral infections peak in the winter, asthma and allergies peak in the spring and fall, and accidents have a summer peak incidence. We provide support for a number of recent assertions in the literature, including: winter peaks in OCD and some kinds of depression, winter onset of Kawasaki disease, and summer increase in urinary tract infection and other bacterial infections. Among the novel findings, we focus on biannual increases in incidence of code 358.01, myasthenia gravis with acute exacerbation (Fig. 1B). This is a potentially serious complication of an autoimmune disease, and thus patients are unlikely to elect to put off a hospital visit. We dissect the cause of this pattern by examining comorbid events occurring before the exacerbation in the medical record, finding that some, notably urinary tract infection, display a seasonal incidence that could help explain this seasonal change in exacerbation.

Fig. 1A

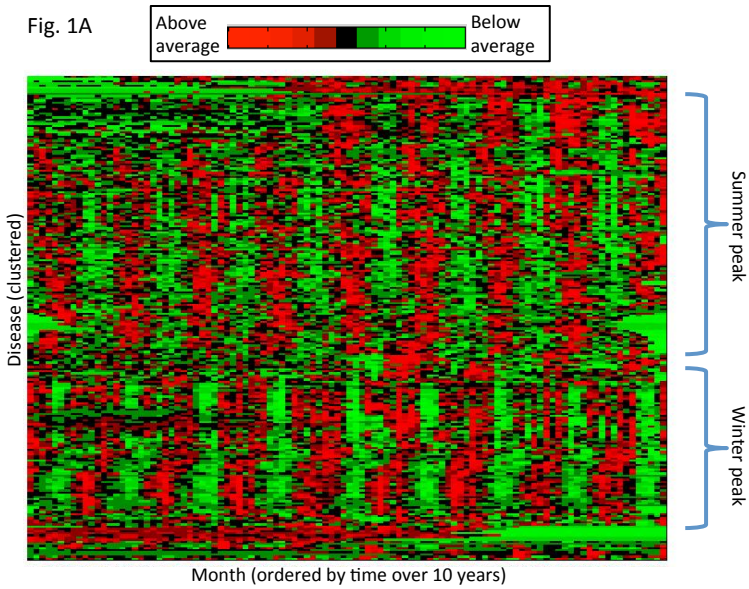


Fig. 1B

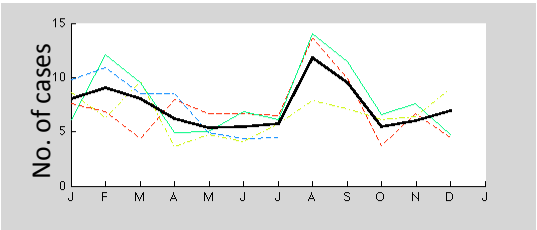


Figure 1A: *Pre-processed and row-normalized monthly incidence for 227 codes with periodic signal*. Each row is a disease, and each column a month over 10 years. Thus, boxes in a row represent incidence of that disease for each month, with red signifying elevated incidence and green decreased incidence. Two main clusters stand out: diseases that occur in the summer (top), and those that occur in winter.

Figure 1B: *Seasonality in monthly incidence of acute exacerbations of myasthenia gravis*. Each of the colored lines represents a year, and the bold black line is the average. Two clear peaks indicate surges in incidence in late winter and late summer.

Protocol Decision Trees to Facilitate Protocol Planning and Patient Enrollment

Joyce C. Niland PhD, Cy Stein MD, PhD, Julie Hom, Adina Londrc MPH, Karen Rickard, Sai Achuthan PhD, Srinivas Bolisetty MS, Mtech, John Meng MD, MS, Ayyappan Nagender, Ajay Shah PhD, City of Hope, Duarte, CA

Introduction

Over the past year, City of Hope (COH) has deployed Protocol Decision Trees (PDTs) covering 40 different oncologic diagnoses and more than 200 protocols to facilitate protocol planning and patient enrollment onto clinical trials. Clinical trial enrollments could be missed when physicians are unaware of available trials that meet their patient's needs. With the PDTs, physicians and research staff are presented with a patient-oriented, eligibility criteria-driven view of all treatment protocols for each cancer. PDTs display in a coherent fashion the paths that can be followed along the decision nodes based on core eligibility criteria, and point out those paths where we do not have an available treatment protocol for a population of patients, suggesting protocol development would be encouraged for this patient profile. It is mandated that PDTs must be consulted to demonstrate that a new protocol being planned will not compete with the same population of patients for which a trial already exists, eliminating overlapping protocols and improving the ability to to be flexible, and studies are added or deleted as they open and close to accrual.

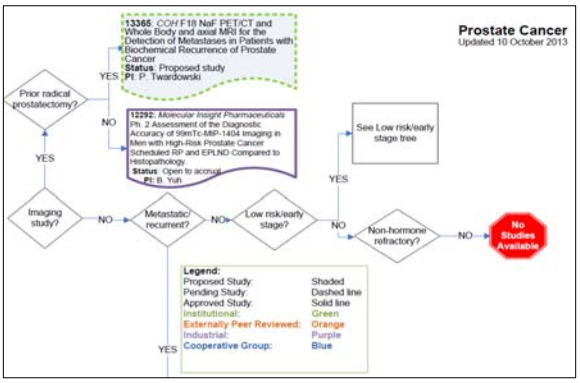


Figure 1: Protocol Decision Tree Example

Methods

Initially PDTs were created and maintained using Microsoft Visio (See Figure 1) to represent the high level eligibility criteria for a given protocol portfolio, by disease. When a researcher initiates a new protocol, he/she works with the decision tree coordinators to include the new protocol within the appropriate PDT. The eligibility criteria are analyzed to determine if there is any overlap with existing protocols. If so, the protocol will not be approved until unique eligibility nodes can be identified to distinguish it from previously approved protocols. Three aspects of this process required informatics support to create a sustainable scalable PDT business process: 1) an automated approach to creating PDTs over Visio diagrams; 2) a branching algorithm to facilitate searching PDTs for available protocols, either when a patient treatment plan is being created, or a new study is being planned; and 3) dynamically populating the PDT with pertinent protocol information from our Clinical Trials Management System (CTMS).

Results

SPIRIT DT (Software Platform for Integrated Research and Transformation Decision Trees module) is used to automate the creation, editing, versioning, and push of the PDTs to the production level on our Clinical Trials On-Line (CTOL) system. Through this process dynamic database queries auto-populate key data elements within the PDT diagrams, including: Protocol Title, Principal Investigator, Sponsor, IRB status, target accrual and patients accrued to date. In addition, an "expert search" functionality has been deployed that branches the user through the most parsimonious set of questions to arrive at a given protocol, or a "stop sign" icon if no protocol is available for the given combination of clinical features representing the core eligibility criteria. This function can be utilized to search for a protocol for a patient presenting in clinic, or for the planning process when an investigator wishes to propose a new trial, to ensure no competing studies already open for the same patient population.

Discussion

As the PDTs are now a mandated component of the COH protocol planning process, and provide an excellent tool for identifying protocols for patients, it became imperative to utilize informatics approaches to render the process scalable, data-driven, automated, and readily extensible. The system eliminates the need to manually enter and update data into the PDTs as it changes. The next goal is to dynamically update the protocol-specific eligibility nodes and automatically add/remove studies as their protocol status changes.

Discrepancy-reducing Feedback Loops Based on Intra- and Inter-Validation of Synoptic Pathology Data

Rebecca A. Ottesen, MS, Leanne Goldstein, DrPH, Kelli K. Olsen, MS, Julie A. Kilburn, BS, Dennis D. Weisenburger, MD, Peiguo Chu, MD, PhD, Joyce C. Niland, PhD, City of Hope, Duarte, CA

Abstract

Synoptic reports are a rich diagnostic data source, yet data errors could adversely affect treatment decisions and research. Internal validation of correlated data fields and cross-validation against Cancer Registry were performed to identify discrepancies. Automated feedback loops reduce errors at origination point and inter-source variability downstream, ensuring highest quality data.

Introduction

The College of American Pathologists (CAP) produces peer-reviewed cancer protocols to standardize data capture of surgical pathology findings. CAP protocol checklists capture key diagnostic fields into consistently structured synoptic reports as a supplement to the traditional unstructured pathology text. Synoptic report data are also a rich information source for secondary use for research. At City of Hope (COH) the synoptic worksheet data are entered via Cerner's CoPathPlus system and housed within our Enterprise Data Warehouse. While this system provides an electronic data capture mechanism, there are no inherent validations or logic checks integrated into the data entry for synoptic reports. Further there is no available data extraction into discrete codified fields, making it necessary to reverse engineer coded data from the final synoptic report. Automated discrepancy-reducing feedback loops are needed to ensure the highest quality data to guide treatment decisions, and for secondary use within research.

Methods

We created the necessary data extraction mapping for breast cancer synoptic report data elements. We then conducted intra- and inter-cross validations of the synoptic data for 1,304 COH breast cancer patients seen from 2007 to 2012. The preliminary internal validation of the synoptic reports consisted of comparing synoptic data elements against correlated fields within the report that should be consistent if accurate. We evaluated 7 such internal consistency logic checks to validate tumor size and lymph node information (e.g. if path 'T3' is entered, then tumor size must be > 5 cm); we also evaluated 5 data fields for missing data. Inter-source cross validations compared synoptic data to Cancer Registry information abstracted by Certified Tumor Registrars into the CNExT relational database directly from the pathology text and synoptic reports. Analyses were conducted using SAS 9.3.

Results

COH began using synoptic reporting in 2007, covering 42% of all breast cancer cases with diagnosis at COH during that time, growing to 87% coverage in 2012. Preliminary results showed that the percent of missing data ranged from 0.1% to 11% across 5 intra- validation checks. Suspect data entries for data capture on tumor size and nodal status across the 7 validation checks ranged from 0.1% to 3%, requiring quality assurance to be conducted. Cross validation of tumor size and nodal information against abstracted Cancer Registry data yielded a range of 7% to 20% in data mismatches. Reasons for inter-data source discrepancies included data entry errors in either the Cancer Registry or the synoptic report, and in limited circumstances special coding rules of the Cancer Registry. Of the errors that arose within the synoptic report itself, we found that often they had been detected and resolved within the Cancer Registry during data abstraction 3-6 months later. Working closely with the Pathology Department, automated feedback loops are being put in place for any intra-source inconsistencies seen within 24 hours post entry, and for inter-source discrepancies found later with the Cancer Registry data abstraction. Results of these feedback loops on data quality over the first 4 months of deployment will be reported.

Conclusion

High quality cancer care relies heavily on valid patient data, as physicians make treatment decisions based on key diagnostic information found in the synoptic reports. National quality metrics in breast cancer care are tied to many of these key clinical variables as well, and high quality data are required for secondary use for research. The synoptic report is important to standardize data capture and promote ease of interpretation; however, the quality of the synoptic data itself is critical for accurate research, analysis and patient care. We predict that automated discrepancy detection with feedback loops to Pathologists will greatly improve this data accuracy.

Impact of Electronic Health Record Alerts on Recruitment for Clinical Trials: A Study of Patients' Perceptions

Emily S. Patterson, PhD[1], Caryn Roth[1], Nancy Elder, MD, MSPH[2], Sian Cotton, PhD[2], Peter J. Embi, MD, MS[1]
[1]The Ohio State University, Columbus, OH;
[2]University of Cincinnati, Cincinnati, OH

Summary

Clinical trial alerts (CTAs) were used to recruit patients for a stroke clinical trial. Structured interviews were conducted with twenty-nine patients. Findings indicate that the CTA was positively viewed and that the participation decision was influenced by physician relationship and the desire to benefit society.

Introduction and Background

Low patient recruitment for clinical trials threatens the validity of study findings. The majority of studies fail to meet their recruitment targets during the originally planned recruitment period. Interventions that use EHRs to increase patient recruitment via Clinical Trial Alerts (CTAs) have been found to be effective in several prior studies. Moreover, in a prior survey of CTA users, the majority of physician respondents (53/69, 77%) appreciated being reminded about an ongoing clinical trial via a CTA. In addition, few respondents felt that the CTA was more than somewhat intrusive (19/69, 27%). The most common reasons cited for routinely ignoring CTAs were lack of time (37%), knowledge of the patient's ineligibility (28%), and limited knowledge about the trial (13%).

Methods

Two EHR-based CTAs were deployed across practices at two institutions. The primary findings from each interview were summarized in four to nine bullet points by a single investigator [EP]. Themes were generated bottom-up and iteratively analyzed from these summaries. The same investigator coded all interviews by gender, whether a patient conducted the interview or a caregiver provided support for the interview, the patient's level of willingness to participate in the stroke trial, and the model of CTA impact on the interaction.

Results

Nine study participants participated in the interviews at the first institution and 20 at the second, for a total of 29 patients. As displayed (Table 1), most patients used the CTA as an insert in shared decision making.

Participate in study?	Direct Info Transfer	Mediated Info Transfer	Shared Decision	No Impact
Yes	0	0	15	4
No	0	2	4	4
	0/0 (0%)	2/29 (7%)	19/29 (66%)	8/29 (28%)

Table 1. Model of CTA Impact for Patients Who Accepted and Declined Stroke Trial Participation

The decision to participate in the stroke study was primarily influenced by the recommendation of a trusted physician and the desire to benefit society in general by participating in medical research. All of the comments about the CTA design were positive, and no negative comments were made with the exception of one minor suggestion for modifying the information about the study on the CTA.

Discussion

Overall, these findings suggest a complex interaction between the CTA, physician, and patient, as opposed to a direct information transfer model. These findings suggest that using CTAs in a direct-to-consumer model, so-called "patient alerts" in Personal Health Records, will be most effective for patients who are strongly motivated to engage in research to benefit society. They are likely to be less effective for patients who are primarily influenced to participate by a recommendation from a trusted provider.

Spreading Research and Engaging Disease Communities – One Automated Tweet at a Time

Katja Reuter[*+1], Ph.D.; Anirvan Chatterjee[*2], B.A.; Bradley Voytek[3], Ph.D.; John Daigre, B.A.[2]
[1]Southern California Clinical and Translational Science Institute (SC CTSI), University of Southern California (USC) and Children's Hospital Los Angeles (CHLA), Los Angeles, CA
[2]Clinical and Translational Science Institute (CTSI), University of California, San Francisco (UCSF), San Francisco, CA
[3]Department of Cognitive Science, University of California, San Diego, San Diego, CA
*Authors contributed equally to the work; [+]Corresponding author (katja.reuter@usc.edu)

Summary
We developed an information system called Science Connect, which uses an automated social media approach on Twitter to disseminate disease-specific research information more widely at little cost. It engaged members of disease communities and contributed to strengthening the research brand of a biomedical research organization.

Introduction
The opportunity: Twitter is a particularly good medium for disease-specific science outreach because of the significant presence of disease communities, as evidenced, for example, by the widespread use of disease-specific hashtags (e.g., #diabetes, #stroke), making it easy for users to categorize and search for disease-related content. In addition, a recent online experiment indicates that more people look at research articles when promoted on social media[1].

The challenge: Research-related information is not easy to find and access for disease communities, and peer-review journal articles and professional presentations are still the two major methods used by researchers to disseminate their work[2].

Science Connect automates the aggregation of disease-specific content based on predefined data sources (e.g., PubMed, ClinicalTrials.gov, university research news), converting content into 140-character messages (tweets) tailored to the microblogging platform Twitter, and automatically posting the tweets at specified times. *Science Connect* also shortens URLs and includes relevant hashtags (e.g., #diabetes, #stroke) to ensure discoverability.

Preliminary Results
We created eight disease-specific Twitter accounts at UCSF (e.g., @UCSFDiabetes). After the initial six weeks, *Science Connect* had distributed a total number of 1,042 tweets across all eight Twitter accounts that were followed by 867 individual Twitter users in total. Across all eight accounts, we found 1,149 clicks on URL links in the tweets and 106 retweets and mentions. We further analyzed the click rate (clicks per tweet) based on different types of content and found that retweets from UCSF university groups, researchers and physicians showed the highest click rate (1.54), followed by research papers from PubMed (1), university research news and clinical trials (0.89), and researchers' profiles (0.64). The data collection and analysis are ongoing and will be completed by the end of 2014.

Conclusion
Our preliminary data indicate that an automated social media approach can be used to distribute research-related information more widely at little cost, that disease communities value such an effort, and that the system helped university groups in charge of communications and science outreach to save time while increasing their information output.

References
1. Terras, MM. The impact of social media on the dissemination of research: results of an experiment. J of Digital Humanities. 2012; 1 (3).
2. Chen PG, Diaz N, Lucas G, Rosenthal MS. Dissemination of Results in Community-Based Participatory Research. m J Prev Med. 2010 Oct;39(4):372-8.

Design and Implementation of an Automated Geocoding Infrastructure for the Duke Medicine Enterprise Data Warehouse

Shelley A. Rusincovitch, Sohayla Pruitt, MS, Rebecca Gray, DPhil, Kevin Li, PhD, Monique L. Anderson, MD, Stephanie W. Brinson, Jeffrey M. Ferranti, MD, MS

Duke Medicine, Durham, North Carolina

Abstract

Geographic information systems (GIS) analysis relies on geocoded data. The Duke Medicine Enterprise Data Warehouse (EDW) has developed an automated infrastructure for geocoding patient address data. Since deployment in August 2012, 4,080,966 patient address records (82.2% of total) have been verified and standardized, and 87.7% of standardized patient address records have been successfully geocoded.

Introduction and Background

Geographic information systems (GIS) seek to combine data from multiple sources (including public health data, census data, and data on the built environment) by geographical location into dynamic maps, location-specific analytics, and geospatial statistics. GIS analysis is widely used in the public health sphere but has only recently been explored at the clinical practice and health system level. This area of research holds great potential for health systems to address population health by combining community-level, environmental, and socioeconomic contextual factors with patient-level clinical characteristics.

Methods

Geocoding—the process of taking textual address information and converting it into geographic latitude/longitude coordinates that can be displayed on a map—is fundamental to all GIS analyses. This process is often semi-manual, iterative, and heavily dependent on geospatial professionals familiar with the specialized software, data, and technology needed to obtain the highest level of geographic accuracy. Recognizing that a semi-manual process for geocoding would not be scalable or sustainable for an enterprise solution, we have developed and deployed an automated process for geocoding patient addresses. Careful consideration was given to maximizing the efficiency of the automated process and attaining high-level geographic accuracy, while keeping patients' protected health information (PHI) secure.

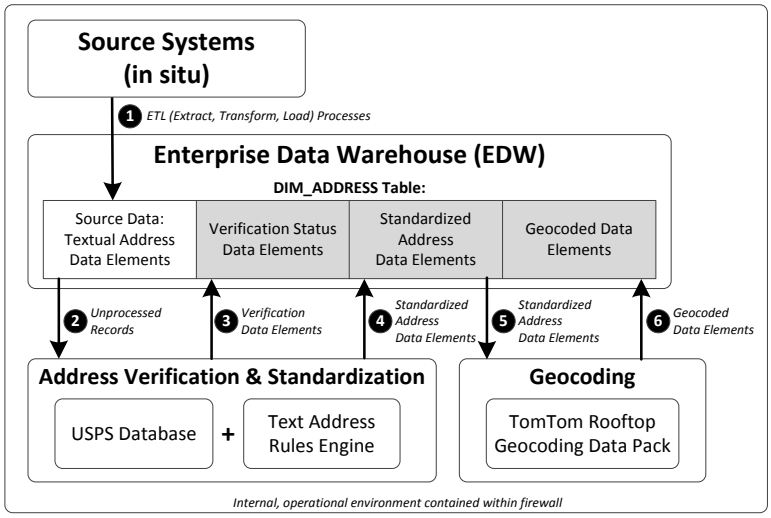


Figure 1. Automated address verification, standardization, and geocoding processes.

Results and Discussion

The automated processes were first deployed in August 2012. To date, 4,080,966 patient address records (82.2% of total) have been verified and standardized, and 87.7% of standardized patient address records have been successfully geocoded. Some source data lack sufficient exactitude or veracity, which results in an inability to verify, standardize, or geocode at a high level of precision; other addresses are not indicative of an actual residence (e.g., a post office box). Data elements that cannot be processed at a sufficient level of accuracy as dictated by our methods are not written back to the EDW; this decision ensures that all data are at a dependable level of precision and quality. As depicted in Figure 1, new addresses in the EDW are processed on a nightly basis. First, the program verifies that the address exists within the United States Postal Service (USPS) database and parses the text string into discrete components, correcting textual inconsistencies. Second, the standardized address data elements are fed into the TomTom rooftop geocoding data pack. Finally, successfully geocoded results are written back to the EDW and incorporated into the address record. The address verification, standardization, and geocoding are performed within Data Management Studio (version 2.1; SAS, Cary, NC, USA). The geocoded data are the basis for a GIS resource that can be accessed, queried, and visualized on demand by Duke clinicians and researchers, regardless of their GIS knowledge and expertise.

Use of the Epic Electronic Health Record for Comprehensive Clinical Research Management at Duke.

Iain Sanderson, BM, BCh, Denise Snyder, MS, RD, Terry Ainsworth, RN, MS, ACNP, Cory Ennis, MS, Julie McCauley, Fabian Stone, MBA, MHA, MT, Leigh Burgess, MHA, MEd, MA, Beth McLendon-Arvik, Pharm.D, Colleen Shannon, JD, and Mark Stacy, MD.
Duke Medicine, Durham, NC.

Abstract: *At Duke University Health System we have implemented the full research functionality of the Epic 2012 Electronic Health Record (EHR) system, building protocols, order sets and study calendars for over 600 active clinical trials at go-live. Combined with a transformation of the operational workflow of clinical research, our implementation challenges assumptions about the role of Clinical Trials Management Systems and their integration with the EHR.*

Introduction: The efficient and cost-effective management of site-based clinical trials are key objectives of the NIH, sponsors and research administrators. Despite this, much of the process of clinical trials management remains spreadsheet and paper-based. A few centers have implemented Clinical Trials Management Systems (CTMS), but very few CTMS have been successfully integrated with the Electronic Health Records (EHR) used for clinical care. This gap is significant, as the typical clinical trials workflow at an AMC involves utilizing the hospital infrastructure for scheduling study appointments, placing research-related orders, documenting evaluations, drawing labs and viewing results. These processes are interwoven with routine clinical care and both the clinical and research revenue cycles, increasing the complexity, and institutional risk, of comprehensive solutions that attempt to integrate clinical trials management and clinical care.

Background: Duke University Health System has deployed the Epic (Verona, WI) EHR, going live successfully across our outpatient and inpatient areas with over 6000 simultaneous users in June 2013. Our research implementation involves the comprehensive management of clinical studies, participants, enrollment, research appointments, study drugs, order sets for research, study billing calendars, documentation of research visits and the management of the research and clinical revenue cycles (split billing).

Methods: "Model" Epic 2012 research functionality was implemented and configured for all clinical trials that invoked our health system's clinical infrastructure, such as those requiring an order, involving a visit in the clinical environment, requiring a lab study or generating a bill. For these studies, Epic has been adopted as the master study registry, master participant registry and the mechanism of managing participant enrollment status. Duke Cancer Institute's clinical trials were built using Epic's Beacon oncology module, with underlying study calendars tailored for each protocol's arm. Our non-oncology implementation deviated from "model" only in our directive to build an order set for each clinical study, enabling a close linkage between orders and the visit cycles of a study calendar. Corresponding study calendars were built with underlying billing codes, payment logic to sponsor or insurance, and with close alignment to a central charge master. Planning for go-live included the building of study administrative records, order sets, study drugs and study calendars for over 600 actively enrolling clinical trials at Duke in a process that engaged study teams, pharmacists, specialized Epic builders, central research administration and the patient revenue management organization in multiple cycles of validation. Over 600 clinical research coordinators received 4 hours of classroom training and an additional 500 research staff participated in online training in the month prior to go-live, with an active communications strategy of town halls, websites, FAQ's, tip sheets, newsletters and "Research Wednesday" events involving the entire clinical research community.

Results: Epic has become Duke's Research Patient Management System with much of the functionality of a traditional CTMS woven into our EHR. This initiative has already forced efficiencies into clinical trials management, from the analysis and agreement on a centrally-managed clinical trials workflow, to the prioritization of studies based on productivity and enrollment, to a rationalization of costs and redundancies inherited in our legacy trial "grids" as they were exposed to the rigor demanded of the new approach.

Discussion: The impending implementation of the Epic EHR at Duke forced a radical assessment of our legacy mechanisms for managing clinical trials, not just in terms of their IT management, but also in terms of the routine operational processes involved in conducting clinical research in our hospital environment and clinics. Of particular concern was a solution to efficiently manage appropriate billing to study sponsor or insurance in the context of a single encounter. A considerable investment was made to configure Epic with an order set or Beacon protocol for each clinical study as we judged this critical to success. However, building Beacon protocols, study drugs, order sets and study calendars in volume, is complex, time consuming and logistically challenging. We believe that this investment, combined with complete institutional alignment, leadership and investigator engagement, has been the key to the success of our "all–in" approach. The integration of traditional CTMS functionality into EMRs is inevitable for complex academic medical centers. Our experience is showing that efficiencies in clinical trial management can be achieved by embracing the evolving functionality in comprehensive EHR systems.

This research was supported the NIH CTSA (Clinical and Translational Science Award) 1UL1TR001117-0.

Government-Developed Software for Clinical Research - Open-Season on Open-Source: The NICHD Clinical Trials Database (CTDB) and Toolkit (CTK) Development and Adoption Strategy

Chandan Sastry, Ph.D.[1], Matthew Breymaier[1], Sean Ivusic, M.S.[1], Asma Idriss, M.S.[1], Thomas P. Caruso, Ph.D., MBA[1,2], Robert Annechiarico[1]

[1]National Institute for Child Health and Human Development (NICHD), National Institutes of Health, Bethesda, MD USA; [2]University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Summary

Over ten years ago, NICHD set out to develop a comprehensive CDMS, the Clinical Trials ToolKit (CTK) primarily to serve NIH Intramural needs, and also making it available to non-profit research institutions. CTK's components, architecture, and integration with other technologies are key to the popularity of the platform with researchers.

Introduction

In recent years, a great deal of controversy has surrounded the development of software by Government agencies and institutes. NICHD, an early entrant into this arena, has been developing and refining the CTK since 2003, rather organically, into a sophisticated suite of clinical research software products, known simply as the Clinical Trials Toolkit (CTK). The program has remained, for the most part, within the NIH's Intramural walls, taking advantage of key elements from other initiatives and standards, such as the cancer Biomedical Informatics Grid (caBIG) as well as other well-known ontologies and libraries. As the platform grew, other institutes in need of such platform system joined the support of the CTK, and now more than a dozen institutes have signed on, as well as several institutions in the extramural world.

Significance

With over 330 active protocols in 13 NIH institutes, as well as several external academic institutions, most of which have contributed a significant body of knowledge and special expertise to the growth of the form and metadata libraries, the collective experience is significant. To-date, investigators and contributing libraries have amassed over 70,000 research questions in the form library, most harmonized to national standards. Although not intended to compete with industry, the recent extramural adoption program for CTK aims at smaller, qualified centers and institutions, that may not have the resources to purchase regulatory compliant (21 CFR Part 11, for instance) platforms, yet play a vital role in the clinical research effort. When compared to other government sponsored initiatives, CTDB stands out in the capability to capture and integrate disparate datasets.

Conclusion

The project has improved data access, security, and research collaboration by calibrating services to an optimal base set of features. This presentation explores the explosive growth of CTK, primarily through the experience of its flagship application, the Clinical Trials Database (CTDB), in support of, but also in contrast to, other NIH initiatives, such as elements of the NCI's caBIG (now NCIP) program, and NCATS' REDCap development. We will explore why CTK has been a success at NIH and why the extramural community is finding this an important tool for support of clinical research efforts, particularly as it relates to the integration of "-omic" data, in support of emerging personalized medicine initiatives.

City of Hope Research Informatics Common Data Elements Information Architecture Framework

Authors: AbdulMalik Shakir, Joyce Niland, Ph.D., Kelli Olsen, Adina Londrc, Susan Pannoni, Stacy Berger City of Hope, Duarte, CA

ABSTRACT

The City of Hope Research Informatics Common Data Elements (RI-CDE) is an application of the information architecture component of the COH Research Informatics Enterprise Architecture Framework (RI-EAF). The RI-CDE is a repository of common data elements, their business and technical metadata, and their semantic relationships. It serves as the foundation for enabling decision support and semantic interoperability within COH and between COH and others.

This poster provides an overview of the City of Hope Research Informatics Common Data Elements conceptual architecture including the standards used, rationale for the choices made, and the anticipated impact on informatics systems interface engineering and procurement activities.

INTRODUCTION AND BACKGROUND

In 2010 the City of Hope (COH) Department of Information Science (DIS) developed an architectural framework for information management entitled "Research Informatics Enterprise Architecture Framework" (RI-EAF).

RI-EAF was developed by the Research Informatics Division of DIS. The purpose of the framework is to provide the standards, guidelines, and procedures required to facilitate the planning, procurement, engineering, and deployment of information systems needed to support research activities at City of Hope.

Research Informatics Common Data Elements (RI-CDE) was initiated in fall 2013. The RI-CDE is an application of the information architecture portion of the RI-EAF. It establishes the methodology by which the semantics of common data elements are harmonized throughout the enterprise to establish a uniform nomenclature for shared concepts and traceability to their realization in information systems, databases, and application interfaces.

The RI-CDE serves as the foundation for enabling decision support and semantic interoperability. In this poster we illustrate its application to defining metadata related to the early capture and encoding of patient diagnosis data.

BUSINESS DRIVERS

Information systems are acquired through three major channels at COH - vendor supplied, internally engineered, and through open source collaborations.

Ensuring semantic consistency among data elements held in common across system boundaries and used in support of divergent business functions, use cases, and user communities is a significant challenge.

The RI-CDE provides the methodology and information architecture conceptual framework for harmonization of the data of particular importance to Research Informatics.

DIAGNOSIS USE CASE

The first Use Case to addressed by RI-CDE is "Patient Diagnosis". Diagnosis is a critical data point for all clinical and research activities; however, there was no organized systemic means for the timely capture and coding of diagnosis data.

Diagnosis data are reliably captured for billing purposes and for mandatory entry in the cancer registry. But each of these processes occurs late in the lifecycle of care for a patient or research subject.

This project helped to highlight the value of capturing a coded diagnosis early in the lifecycle of care such as during new patient services, patient scheduling, and patient registration activities. It also highlighted heretofore unrecognized disparities in the definition and encoding of diagnosis data.

STANDARDS USED

RI-EAF is based, in part, upon The Open Group Architecture Framework (TOGAF)[1]; the Health Level Seven (HL7) Services Aware Interoperability Framework (SAIF)[2]; the HL7 Common Terminology Services Release 2 (CTS II)[3]; and the ISO 11179-3[4] Metadata Registry Meta-model.

References

[1] http://www.opengroup.org/togaf/

[2] http://wiki.hl7.org/index.php?title=Product_SAIF

[3] http://www.hl7.org/documentcenter/public/standards/dstu/2009may/V3_CTS_R2_DSTU_2009OCT.pdf

[4] http://metadata-stds.org/11179/

Standards-based Representation of Open Clinical Trials Data for Public Dissemination and Reanalysis

Ravi D. Shankar, MS[1], Atul J. Butte, MD, PhD[1]
[1]Division of Systems Medicine, Stanford University, Stanford, CA

Abstract

Public access to clinical trials data has created tremendous opportunity to perform meta-analysis of clinical trials. We are developing an analytical framework that employs standards-based clinical trial ontologies. We will use this standard to disseminate raw individual-level clinical trials data through the NIAID sponsored ImmPort data repository.

Introduction

The demand for broad open access to entire clinical trials data is on the rise. This access to raw clinical trials data extends beyond the summary form generally available via clinical trial publications or public clinical trial metadata repositories such as ClinicalTrals.gov. Public access to clinical trials data can create tremendous opportunity to perform cross analysis of clinical trials and to combine clinical trials data with other biological and environmental datasets, in order to evaluate new research hypotheses that were not originally formulated in the studies. But such analysis of disparate data presupposes a) uniform representation of clinical trials data using data standards, and b) easy access to such standard representations of clinical trial data in analytical environments.

The Dissemination and Reanalysis Framework

The Immunology Database and Analysis Portal (ImmPort) system warehouses clinical trials data in all areas of immunology that is generated by scientific researchers supported by the National Institute of Allergy and Infectious Diseases (NIAID). Within ImmPort, we are developing an analytical framework that employs standards based clinical trial ontologies to support meta-analysis of clinical trials. We have targeted the open-source R statistical environment in our initial implementation. Our framework (Figure 1) comprises of four main components: 1) a clinical trial ontology - a formal specification that encapsulates clinical trial data, 2) an authoring tool that enables users to easily create clinical trial knowledge bases by encoding specific clinical trials using the clinical trial ontology, 3) an R class generator that creates clinical trial R classes and properties that correspond to the ontology's classes and properties, and 4) an R object generator to create

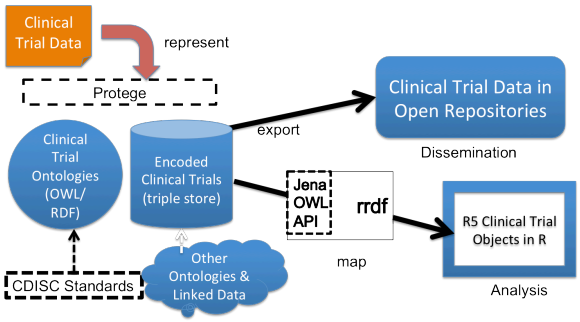


Figure 1: Components of the framework to support public dissemination and reanalysis of clinical trial data

clinical trial R objects by populating the R classes with clinical trial knowledge. The ontology leverages CDISC standards such as PRM, SDTM and BRIDG, and incorporates terms and semantics found in these standards. We have used semantic web technologies such as OWL/RDF to specify the ontology and to encode trials using the ontology. The ontology represents clinical trial entities including study design, outcomes and endpoints, planned activities, study subjects, clinical and experiment data. Specific clinical trials are encoded using the ontology as RDF triples and stored in an RDF triple store. Using Jena OWL API and RDF SPARQL queries, the R clinical trial objects are populated with the clinical trial knowledge in the triple store. Thus, the clinical trial data is readily accessible in R for analysis, already working as of this writing as a prototype.

Conclusion

By standardizing the representation and access of clinical trials data in the R environment, our framework supports efficient cross analysis of clinical trial data using R. We are building and evaluating our framework with real data from clinical trials currently hosted in ImmPort. By basing our work on formalisms such as BRIDG and CDISC standards, we ensure that our framework can be used in other domains, while extending the use of these two important standards frameworks into the new direction of open science.

ClinMiner: Ontology Based Clinical Data Portal

[1, 2]Mary Shimoyama, PhD; [2]Tomasz Adamusiak, PhD

[1]Department of Surgery, Medical College of Wisconsin, Milwaukee, Wisconsin; [2]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee Wisconsin

Summary Few applications are designed to accommodate data imported from the EHR as well as data generated as part of a research protocol. ClinMiner is an ontology-based clinical data portal that provides the functionality to accommodate the varied needs of projects such as treatment effectiveness analysis, clinical sequencing, and epidemiological studies.

Introduction Standardizing data from multiple sources is a common problem for clinical researchers who integrate data from the electronic medical record with data generated as part of their studies. Differing formats, labels and notations create barriers to true integration. Ontologies have been widely used to standardize metadata tags and simple biological annotations, and their utility in defining standard data formats for clinical data integration has started to become more common. Described here is an application with the flexibility to be customized for many types of clinical research.

Methods ClinMiner provides a multi-component platform consisting of a relational database, data entry and curation software, cohort discovery tool and annotation report tools providing users the ability to customize reports on single or multiple patients or subjects with a download option. Data can be entered manually or uploaded from clinical data warehouses or other databases. The ontologies of Meaningful Use – SNOMED CT, LOINC and RxNorm provide the framework for data formats and standardize information within the frameworks for demographic information, diagnoses and symptoms, medications, laboratory test results and procedures. ClinMiner provides a UMLS based ontology browser to assist users in determining correct annotations for manual data entry and to provide links between familiar vocabularies used in other applications such as ICD 9 and CPT. Ontology import pipelines and mapping techniques allow for additions of new and updated ontologies with mappings to existing ontologies and previous versions of ontologies. A query wizard for cohort discovery and a reporting system allows users to customize the return of desired information on identified subjects and provide a cross participant analysis. Individual participant results present summary data, a chronological flow chart view, a timeline view and a chart view which presents vitals and laboratory measurements across time.

Results and Discussion Data has been successfully integrated from a variety of sources including the Clinical Avatar Project which contains 100,000 created patients, the Clinical Data Warehouse at the Medical College of Wisconsin (MCW) which houses data extracted from the EPIC Electronic Medical Record system used at its associated hospital and data extracted from paper documents as part of the clinical sequencing project at MCW. ClinMiner provides a complete platform for a variety of research projects and easy customization through its ontology-based design. Continued expansion of the platform will include customized data downloads, additional data visualizations across participants and links to statistical analysis and other software tools.

When Should We Share? Securely Measuring the Overlap Between Private Datasets

S. Joshua Swamidass, MD PhD[1*], Matthew Matlock[1], Leon Rozenblit PhD[2]
[1]Department of Pathology, Washington University School of Medicine, St. Louis, MO;
[2]Prometheus Research LLC, New Haven, CT

Abstract

Some of the most interesting and powerful datasets---like health records, genetic data, and drug discovery data---cannot be freely shared because they contain sensitive information. In many situations, knowing if private datasets overlap determines if it is worthwhile to navigate the institutional, ethical, and legal barriers that govern access to sensitive, private data. Here, we report the first secure method of publicly measuring the overlap between private datasets.

Introduction

Integrating and analyzing large amounts of data is proving to be a powerful method for exploring and understanding everything around us. In an increasingly data-driven world, we have two contradictory impulses, both to *share* data and to keep it *private*. On one hand, data is powerful and sharing it lets us answer big questions that are unapproachable by other means. On the other hand, there is real danger in sharing it publicly or indiscriminately, so there are institutional barriers governing access to useful data.

Often, especially when human subjects are involved, the value of sharing data depends on the overlap between private datasets. The amount of overlap determines if there is reason to work through institutional, legal, or ethical barriers governing access to private data. Herein lies the problem, we often need to know the overlap between datasets to justify sharing. But how can we know the overlap between private datasets before sharing them?

We solve this problem with a new, secure way of measuring the overlap between two private datasets (Figure 1). This method uses a public algorithm to generate a public summary of any private dataset's contents, its cryptoset. The overlap between two private datasets can be estimated by comparing their cryptosets. At the same time, it is not possible to determine which items are in a private dataset from its cryptoset. Unlike other approaches to this problem the item-level security arises from statistical properties of cryptosets rather than the secrecy of the algorithm or computation difficulty, so cryptosets can be shared in public, untrusted environments.

Conclusion

Cryptosets are informative about overlaps between private datasets with a stable, tunable accuracy. For the foreseeable future, the conflict between our desire to share data and our need to protect it will continue. However, sharing private data is increasingly necessary for scientific progress in fields that are dominated by sensitive information. Cryptosets may be informative and secure enough to help navigate barriers to accessing data, and enable collaborations and studies not otherwise possible.

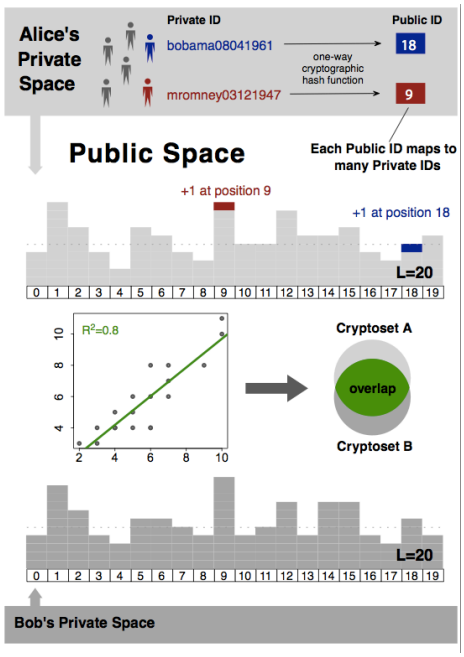


Figure 1. Cryptosets are shareable summaries of private data, from which estimates of overlap can be computed. They are constructed using a cryptographic hash function to transform private IDs from a dataset into a limited number of public IDs, and then combining these public IDs into a histogram. From this histogram (about 1000 IDs long in practice), the overlap between private datasets can be estimated in a public space. The security of cryptosets relies on the fact that several private IDs map to each public ID. The estimates are based on the Pearson correlation between cryptosets, and can only measure overlap at a predetermined resolution.

RexInstrument: Exploring an Open-Source Standard for Configuring Clinical Research Instruments

Charles Tirrell, BS, Leon Rozenblit, JD, PhD, Frank Farach, PhD
Prometheus Research LLC, New Haven, CT

Summary: Electronic data capture and sharing are critical to clinical and behavioral research but are hindered by the lack of an open-source standard for configuring clinical assessments electronically. We describe the development of a prototype standard that allows us to configure an instrument once for delivery via multiple technologies.

Background: Configuring a clinical or behavioral research instrument to be used in an electronic data capture (EDC) system requires the specification of metadata describing the content, structure, display, and presentation logic for the instrument form. Many EDC systems collect these metadata elements in an application-specific manner, making it difficult to share and repurpose instrument configurations for different EDC technologies. These barriers could be removed by the adoption of an open-source, application-neutral instrument configuration standard. Ideally, such a standard would (a) define all instrument metadata needed for delivery (e.g., questions, field data types, choice lists, show/hide/disable logic); (b) support the transformation of the instrument definition file into formats supported by a variety of platforms (e.g., web browsers, tablets, and mobile devices); (c) be human readable and editable; (d) be extensible with respect to unanticipated attributes, such as those needed for delivery via SMS or voice prompt systems; (e) be available under a liberal open-source license that would encourage the evolution of a user/developer community; and (f) encourage the creation of an open-source, revision-controlled repository of research instruments hosted online (e.g., on Bitbucket or GitHub). Unfortunately, no existing standard meets all of these criteria. REDCap's commonly-used configuration format is not open source and cannot be compiled into any format other than REDCap; CDISC's SHARE, while leveraging the powerful Object Definition Model (ODM), is a non-open-source commercial service. We describe our first attempt to create such a prototype standard (RexInstrument) below.

Methods: Prometheus has developed four different EDC systems for clinical research between 2003 and 2013, and our current implementation of our flagship open-source product, the Research Exchange Database (RexDB), contains two different EDC technologies, one (RexEntry) optimized for staff data entry, the other (RexAcquire) for self-administered questionnaires. The former has presentation requirements that optimize for speed and complex skip logic. The later optimizes for ease of use, cross-platform consistency, and good help screens. Both face complex challenges with instrument versioning and internationalization of instrument content. Importantly, both share many definition elements (question text, hints, pagination, etc). Further, data consumers require the data resulting from both cases to be merged together into a single data exploration source, implying a unified definition model. Because of our unusual experience in developing multiple EDC systems, we were able to review and compare several sets of requirements and constraints in an attempt to derive a unified, extensible definition model that would allow us to compile instruments from configuration files across multiple delivery methods.

Results: Our review focused on four areas for core instrument configuration data: (1) data element definition (variable name, data type), (2) rapid data entry parameters (keyboard shortcuts, show-when logic, skip logic), (3) question-level display information (question titles, hints, hide/disable logic), (4) grouping information (pagination, sections), and (5) mode-specific presentation instructions. The analysis allowed us to derive a unified instrument definition that allows us to compile for both RexAcquire and RexEntry from a single instrument configuration file. To achieve this, we made preliminary decisions about what elements should be universal for all EDCs, and what configuration elements must be extensible. We also settled on a common data serialization standard for the configuration files. After considering XML, YAML, and JSON, we settled on JSON, since it is an increasingly dominant data interchange format in both web and mobile technologies.

Discussion: Our early finding are encouraging, and suggest that an open-source unified instrument definition standard can be developed with modest effort to cover the common 80% of the cases. Some edge cases that may require significant additional effort have already emerged. For example, we haven't reached a conclusion on how support internationalization, partly because our most pressing immediate use-case requires support for right-to-left languages; this presented some unexpected challenges with dynamically modifying screen layouts. Other work that remains is exploring the ability to translate from and to the REDCap instrument definition standard and from and to the CDISC ODM. We plan to deliver tools for two-way REDCap translation as part of the open-source project, along with a call to submit draft instruments to an open revision control system under appropriate licenses. CDISC translation will require considerable analysis and may be better taken on as an open-source collaboration across multiple institutions. Emergence of an open-source standard for instrument configuration will streamline EDC and encourage the development of a variety of tools suitable for delivery of content under different use conditions. Further, development of a shared open-source revision-controlled instrument library based on an open standard will enable better data sharing and interoperability across research programs and institutions. The need of an open-source instrument definition standard and a library of instruments based on that standard may be obvious, but these ideas have not yet been implemented; they should be.

Understanding Diagnosis Assignment from Billing Systems Relative to Electronic Health Records for Clinical Research Cohort Identification

Lemuel R. Waitman[1], Kelly Gerard[2], Daniel W. Connolly[1], Gregory A. Ator[3]
[1]Division of Medical Informatics, Department of Internal Medicine; [2]Center for Health Informatics; [3]Department of Otolaryngology, University of Kansas Hospital, University of Kansas Physicians; University of Kansas Medical Center, Kansas City, Kansas

i2b2 is used to explore agreement between billing records, the EHR, lab results, and quality measures for acute myocardial infarction. Having multiple data sources supporting the diagnosis enhances cohort characterization.

Introduction: Integrated data repositories, allow investigators to access patient data across electronic health records (EHR), billing systems, and quality/research registries but with the burden of making informed choices from this richer picture of longitudinal health. Prior studies have compared billing records against chart review "gold standards", but more automated comparisons between billing, EHR and other integrated data sources is critical to:

1. Researchers defining cohorts for retrospective analysis, study feasibility, or trial recruitment.
2. Understanding how reliably different features of the EHR are used to document diagnoses.
3. Motivate methods for understanding the disease progression and how multiple sources of diagnosis – during, before, and after the encounter – may provide increased support for defining the research cohort.

Methods/Results: Using acute myocardial infarction (AMI) and congestive heart failure (CHF), we evaluated the agreement between cohort definitions using HERON, an i2b2 based integrated data repository that incorporates the EHR (Epic), the University Healthsystem Clinical Database containing quality measures and hospital billing codes, and an ambulatory billing system (GE IDX) used by university physician clinics. Different sources for diagnosis are represented using i2b2 modifiers. Table 1 shows AMI hospital billing diagnosis intersected with 1) different segments of the EHR; 2) the broader class of ischemic heart disease in the EHR; 3) an abnormal cardiac marker (tropoinin) result; 4) ambulatory billing records; and 5) an abstracted quality measure for AMI. Agreement is reported when during the same encounter as well as independently across the patients' integrated records.

Table 1: Number of patients with acute myocardial infarction diagnoses assignment by hospital billing records relative to other sources of diagnoses code assignment (percentages relative to hospital diagnosis).

	Hospital Billing Diagnosis 1,367		**Primary Billing Diagnosis 920**	
	Independent	***Same Encounter***	***Independent***	***Same Encounter***
Clinical EHR Diagnosis, any source, 5,520	1,136 (83%)	1,058 (77%)	890 (97%)	801 (88%)
Encounter diagnosis 2,111	1027 (75%)	976 (71%)	838 (91%)	770 (84%)
Hospital problem 55	30 (2%)	22 (2%)	21 (2%)	11 (1%)
Medical History dx 4092	415 (30%)	0	334 (36%)	0
Primary diagnosis 296	135 (10%)	42 (3%)	118 (13%)	38 (4%)
Principal problem 16	9 (1%)	9 (1%)	9 (1%)	8 (1%)
Problem List 1938	1,011 (74%)	731 (53%)	807 (88%)	520 (57%)
Ischemic Heart Disease (ICD9 to 410-414.99) 29,991	1,294 (95%)	1,248 (91%)	910 (99%)	896 (97%)
High Troponin I, or Point of Care > 0.05ng/mL 8,169	1,302 (95%)	1,287 (94%)	891 (97%)	888 (97%)
Clinic Billing Diagnosis (IDX) 825	573 (42%)	na	476 (52%)	na
AHRQ Quality In-hospital mortality AMI 910	910 (67%)	910 (67%)	910 (99%)	910 (99%)

Discussion: EMR features to identify hospital and principal problems are unreliably recorded. The problem list was updated only 57% of the time for hospitalizations with AMI as the primary diagnosis (44% for CHF). As expected, AMI wasn't recorded during the encounter as past medical history but was in past medical history approximately one third of the time. That may be due to follow up by other practices. Future work will seek to develop methods that evaluate diagnosis assignment before and after the encounter to understand how acute diagnoses are subsequently reported as past medical history; supporting continuity of care. Such methods may also provide feedback to clinical leadership concerned with clinical documentation quality and timeliness.

[1]Kiyota Y, Schneeweiss S, Glynn RJ, Cannuscio CC, Avorn J, Solomon DH. Accuracy of Medicare claims-based diagnosis of acute myocardial infarction: estimating positive predictive value on the basis of review of hospital records. Am Heart J. 2004 Jul;148(1):99-104.

Components and Workflow for Patient Identification Using i2b2 for Clinical Trials (i2b2-CT)

Nich Wattanasin, MS[1], Michael Mendis[1], Alyssa J. Porter, MS[1], Stella Ubaha[1], Jonathan Bickel, MD[3], Kenneth D. Mandl, MD, MPH[3], Isaac S. Kohane, MD, PhD[3], Shawn N. Murphy, MD, PhD[1, 2]
[1]Partners HealthCare, Boston, MA; [2]Massachusetts General Hospital, Boston, MA; [3]Boston Children's Hospital, Boston, MA

Abstract

The recruitment of a sufficient number of patients is a crucial part of clinical trials research. We have developed components and leveraged community projects in i2b2, an open source platform for secondary-use of clinical data for research, to accelerate the process of identifying and reviewing suitable patients for a study.

Background

Informatics for Integrating Biology and the Bedside (i2b2) is one of the sponsored initiatives of the NIH Roadmap National Center for Biomedical Computing. The primary goal of i2b2 is to provide clinical investigators with a cohesive set of software tools necessary to collect, host, and manage clinical data from the EMR, and enable the secondary-use of that data for research. The i2b2 platform is designed in a modular fashion consisting of interoperable "cells", or software modules that communicate through web services, which fosters the invention of additional plugins contributed by the i2b2 community that extends the functionality to a new "i2b2-CT" platform, which supports end-to-end identification of patients for clinical trials research. Today, within the i2b2 web client, an investigator can render views of patient centric data for review, as well as utilize newly developed tools, such as our clinical trial suite of apps, to make the review process more efficient for identifying patients qualifying for clinical trials.

Methods

We have developed a number of new components to enhance the i2b2 platform to support the patient recruitment process for clinical trials research. These elements include 1) a web service based ETL cell to retrieve the latest patient data from an institution's enterprise web services, 2) an identity management cell to manage encrypted patient lists and authorizations, and 3) web client plugins to support a tabular display for patient selection. In addition, we leverage community-driven projects such as the SMART-i2b2 platform (Substitutable Medical Apps Reusable Technologies) to further augment the determination of trial suitability workflow and allow for investigators to view PHI in a customizable patient-centric view. Utilizing the SMART application programming interface (API), we designed and developed a suite of clinical trials related SMART apps that run inside the i2b2 web client and allows one to manually specify eligibility criteria for a clinical trial or automatically import the criteria from ClinicalTrials.gov. The apps comb over the patient's data on a patient-by-patient basis, for example, looking for a certain medication or problem, or a text string in a note, and displays a matching score based on the defined criteria, allowing the investigator to flag the patient as a potential match.

Results

We have implemented and deployed a limited release of the work described herein at Partners HealthCare for select i2b2 disease-based driving biology projects. The goal is to deploy i2b2-CT at both Partners HealthCare and Boston Children's Hospital by target date mid-2014. All of the new components that we have developed and community-contributed work previously mentioned are open source and available for download on the i2b2 community wiki site at http://community.i2b2.org

Conclusion

Implementing an end-to-end patient identification workflow for clinical trials in a robust and extensible framework such as i2b2 allowed us to not only leverage existing community-driven projects, but also look forward to emerging initiatives based on this groundwork. One such project is the Shared Health Research Informatics Network (SHRINE) for distributed i2b2 queries which could support recruitment for multi-site clinical trials. This work was sponsored by Harvard CTSA, NIH U54LM00874 and ONC 90TR0001/01.

Research networking across an entire university

Griffin M Weber, MD, PhD
Information Technology and Center of Biomedical Informatics
Harvard Medical School; Boston, MA, USA.

Summary: Harvard Faculty Finder (HFF) (http://facultyfinder.harvard.edu) is a new website that uses the open source Profiles Research Networking Software (RNS) platform to create a university-wide view of all 11,000 Harvard faculty and their scholarship. This presentation describes the challenges in building HFF and suggestions for others building university-wide research networking tools.

Introduction: We built HFF to help students, faculty, administrators, and the general public locate Harvard faculty according to research and teaching expertise. It brings together schools of Arts & Sciences, Business, Dentistry, Design, Divinity, Education, Engineering, Government, Law, Medicine, and Public Health.

Methods: We faced several challenges in developing HFF, and our approach to these can serve as a model for other institutions looking to extend research networking tools beyond biomedical faculty to their whole university:

- Name disambiguation: HFF imports data from numerous sources, including publications from Thomson Reuters' Web of Science (WoS), patents from the US Patents and Trademark Office, teaching from the Harvard Course Catalog, and books from the Harvard Library. The difficulty is matching an item to the right person, especially with common author names like "J Smith". We started with a simple point-based disambiguation algorithm (e.g., a last name match gets 50 points, a first name match gets 30 points, etc.), which matches content to faculty if the points add up to a particular threshold. For each data source, we then adjusted the algorithm for the type of metadata available (e.g, some books have the author's year of birth, and patents have the inventor's state and city).
- Subject hierarchies: Although numerous domain-specific vocabularies exist (e.g., MeSH in biomedicine), there isn't one standard ontology that covers all disciplines. Because most faculty have at least one WoS article (unlike other content types, such as patents, which only a few faculty have), we used the ~250 WoS Subject Areas as the top level in the hierarchy. To obtain a next level down, we apply a Term Frequency Inverse Document Frequency (TF-IDF) algorithm to the keywords associated with all WoS articles of a given Subject Area. By manually mapping WoS Subject Areas to the Classification of Instructional Programs (CIP), we could place courses taught by Harvard faculty into this subject hierarchy.
- Gaining support: In order to gain university-wide support for HFF, we had to balance the schools' desire to connect databases across Harvard with their reluctance to create a central faculty "profiling" website, which might compete with schools' own faculty websites. Thus, the HFF website focuses on *searching* for faculty, but it does not display a full profile of each person. Instead, it makes the complete data about faculty available only through web services and RDF, which other systems can use to display the information. In other words, we had to *remove* functionality from the open source version of the software in order to get buy-in from the schools. We also had to adjust parameters in the search algorithms. There were concerns that the defaults in the open source software would unfairly rank biomedical faculty higher because of the large number of articles they publish. We therefore decreased the importance of a journal article relative to books, courses, and other content types.

Results: After being restricted to the Harvard community during an initial testing phase, HFF was made available to the public in December, 2013. Although license restrictions might limit other institutions from using WoS in the same way as HFF, we added a new feature to the open source Profiles RNS software that imports publication data purchased through Elsevier's Scopus Author Refinement Service. The HFF disambiguation algorithms for patents and grants will be incorporated into an upcoming release of the open source software. The subject hierarchies are available in the Browse feature of the public HFF website. The specific features of the open source software that we turned on or off in HFF and the parameter values we used for the search algorithm were based on the politics and culture of Harvard; however, other universities can learn from this to customize the settings in their own websites.

Discussion: The HFF website is available at http://facultyfinder.harvard.edu. The open source Profiles RNS code can be downloaded from http://profiles.catalyst.harvard.edu. For the first time, HFF has created a university-wide view of faculty scholarship at Harvard.

Are EHR Data Suitable for Secondary Use? Researcher Views

Nicole G. Weiskopf, MA, Suzanne Bakken, RN, PhD, Chunhua Weng, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

The secondary use of electronic health record (EHR) data has the potential to improve the efficiency and representativeness of clinical research. Concerns about data quality and suitability, however, serve as a limiting factor in the reuse of EHR data. In order to understand these concerns and inform future secondary use research efforts, we conducted eleven semi-structured interviews with researchers in the Columbia University Medical Center (CUMC) community. Our findings indicate significant reservations about the quality of EHR data, as well as the belief that the potential value of these data nevertheless makes their reuse an important goal. Researcher perceptions of EHR data quality should inform efforts to address approaches to data quality assessment.

Introduction

The EHR data contained in clinical data warehouses and repositories represent an opportunity for retrospective research. Nevertheless, there are significant concerns amongst researchers and clinicians regarding the quality of EHR data and the impact of these quality issues on the data's suitability for secondary use. In order to enable the reuse of EHR data, it is necessary to understand these concerns from the point of view of potential data consumers. To our knowledge, however, researcher views on this topic have not previously been studied. Therefore, we endeavored to gain an understanding of clinical researchers' perceptions of EHR data quality and their attitudes towards the suitability of EHR data for secondary use.

Methods

Included in this analysis were nine semi-structured interviews with clinical researchers in the CUMC community. Three broad topics were addressed: experiences using EHR data, perceptions of EHR data quality, and intentions to reuse EHR data. The Columbia University Health Sciences IRB approved this study. Participants were identified through purposive sampling and selected to reflect a range of backgrounds. The interviews were analyzed using descriptive content analysis with a combined inductive and deductive approach informed by the Precede-Proceed Model, specifically the concepts of predisposing, enabling, and reinforcing factors.[1] Validity was supported through member checking and the use of an audit trail.

Results

Respondents had significant concerns about the quality of EHR data, but all were either currently using EHR data in their research or wished to do so in the future. Factors that predisposed the participants towards or against the reuse of EHR data include their beliefs about EHR data quality, the suitability of the data for research, and data verification processes that should take place before reusing EHR data. Most believed that a prerequisite of reuse was review by someone familiar with the data and the process by which the data were recorded in the EHR, preferably a clinician. The primary categories of data quality identified by the participants were completeness, correctness, concordance, clinical granularity, fragmentation, structuredness, and signal-to-noise. Major challenges making it difficult to enable the reuse of EHR data included the limited accessibility of the data and the difficulty of identifying relevant cohorts. The desire to reuse EHR data was reinforced by the promise of accelerating medical research and publication.

Discussion

Although most of the researchers interviewed had embraced the secondary use of EHR data, all expressed concerns about data quality. The desire for review of the data by clinicians prior to reuse makes the secondary use of EHR data not only a time-consuming process, but also a potentially unsustainable one if we wish to move towards the use of de-identified datasets. Moreover, it would be difficult to judge whether or not clinician review of EHR data would be sufficient for ensuring data quality and validity of research results. Overall, we have gained an understanding of the specific concerns that researchers have in regards to the secondary use of EHR data. This information should assist us in informing future efforts to enable the reuse of EHR data in research.

References

1. Green, LW, Kreuter, MW. Health Program Planning: An Educational and Ecological Approach. 4th ed. NY: McGraw-Hill Higher Education; 2005.

Comparison of Medication Extraction Methods in the Cleveland Clinic Electronic Health Record.

Brian J. Wells, MD, PhD[1]; Alex Milinovich, BA[1]; Sandra Griffith, PhD[1]

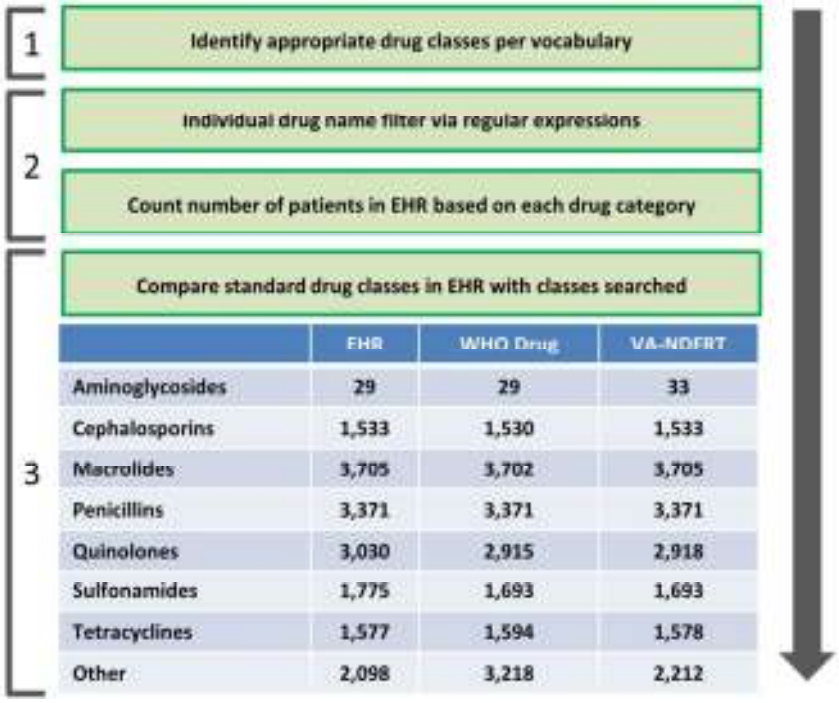
1. Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH

Summary: This study compared methods for extracting antibiotic prescriptions from the EHR at the Cleveland Clinic. Regular expression searches of prescriptions using antibiotics compiled from the VA-NDFRT and the WHO Drug DDE showed similar results compared to the default classification except that the VA-NDFRT method overestimated the number of patients receiving "other" antibiotics.

Background: Research using electronic health record data frequently requires classification of patients according to therapeutic classes of medications. The default medication classification scheme in the Clarity database of the Epic EHR at the Cleveland Clinic sometimes has missing or incorrect information. Manually compiling lists of medications is time consuming and requires frequent updating. The purpose of this study was to compare different methods for medication identification.

Methods: This study obtained all outpatient, oral prescriptions ordered at the Cleveland Clinic between 12/1/2012 and 12/15/2012. The default classifications in the EHR were used to identify the number of unique patients receiving at least one prescription in each of the antibiotic therapeutic classes listed in figure 1. In addition, medication lists for these classes of antibiotics were also compiled from the VA-NDFRT and the WHO DRUG DDE. Regular expressions were used to search for prescriptions matching any of the drug names from these databases in each of the respective antibiotic categories.

Results: Therapeutic class designation was missing in less than 1% of all prescriptions in the EHR after limiting medications to oral products. Table 1 shows the number of unique patients who received at least 1 prescription in each of the drug classes. Overall the results were quite similar between the three methods. However, the Who-Drug method captured a substantially larger number of patients receiving "other" antibiotics when compared to the other methods.



	EHR	WHO Drug	VA-NDFRT
Aminoglycosides	29	29	33
Cephalosporins	1,533	1,530	1,533
Macrolides	3,705	3,702	3,705
Penicillins	3,371	3,371	3,371
Quinolones	3,030	2,915	2,918
Sulfonamides	1,775	1,693	1,693
Tetracyclines	1,577	1,594	1,578
Other	2,098	3,218	2,212

Discussion: The similarity of the results was surprising to our team, detailed comparisons of discordant patients is necessary to determine the exact cause of these discrepancies. Additional studies are also indicated to determine if similar results are obtained with other medication types (e.g. hypertensive medications). It may be possible to rely on the default classification schemes for most projects without causing substantially different research results.

Temporal Knowledge Acquisition from Clinical Research Documents for Community-based Clinical Research Data Standards Development

Chunhua Weng, PhD
Department of Biomedical Informatics, Columbia University, New York City

Abstract

The biomedical informatics community is increasingly baffled by a dilemma, "a sea of standards: sink or swim?" To address this important problem, this presentation will (1) review existing methods for standards development and discuss opportunities for improvement; (2) propose a new approach to acquiring knowledge of "folk standards" that have been in use, as reflected in clinical research documents, to inform official clinical research data standards development; and (3) demonstrate the usefulness of this approach by scanning ClinicalTrials.gov to identify frequent eligibility features for various cancer clinical trials between 1999 and 2013.

Introduction

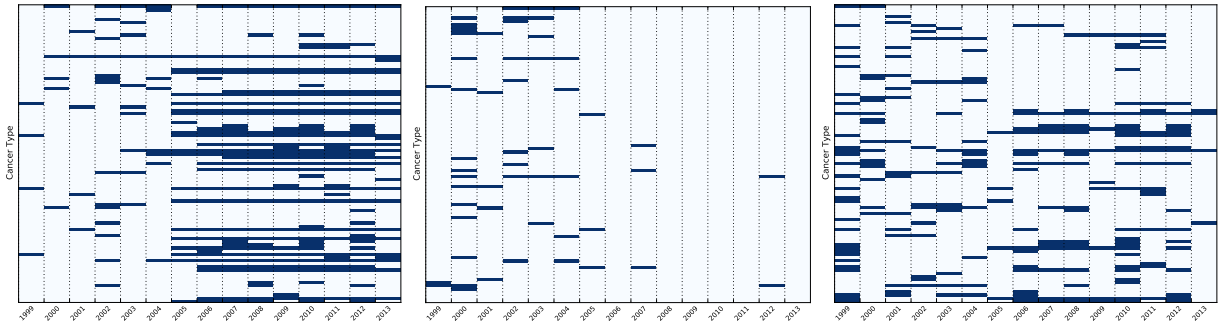
The field of biomedical informatics is increasingly suffering from the tension between numerous clinical research information standards[1], available or under development, and sparse adoption of these standards by clinical researchers and vendors. To supplement the current top-down model of "standards development by experts followed by dissemination to users", we propose a temporal knowledge acquisition method for uncovering frequently used data elements in clinical research text to facilitate community-based knowledge sharing and bottom-up standards development. We hypothesize that this method can identify popular content used frequently in the research community and thereby increase the adoption of standards derived "from researchers and for researchers".

Methods

Using the free-text trial summaries on ClinicalTrials.gov, we applied an unsupervised tag mining method[2] to identify 929 frequent eligibility features that each appear in at least 5% of 40,204 clinical trials for 95 cancer types. Then we analyzed the temporal patterns between years 1999 and 2013 for networks of cancer types and frequent eligibility features to infer cross- and within-specialty knowledge sharing patterns among cancer clinical trials over time.

Results

The figures below display the temporal usage of three example frequent eligibility features: (1) hypersensitivity; (2) creatinine; and (3) creatinine clearance, respectively. Their y-axis indicates 95 cancer types and x-axis indicate 15 years (1999-2013). Each blue marker indicates that the eligibility feature appears in at least 5% of the trials for the corresponding cancer type for the corresponding year. Since 2005, "hypersensitivity to chemotherapy" has been adopted by clinical trials for up to 41% cancer types as a frequent eligibility feature, while previously its usage ranged between 5% and 10%. Meanwhile, starting in 2006, creatinine clearance has gradually replaced creatinine to indicate kidney function for the clinical trials of between 18% and 27% of cancer types.



Conclusions

Our approach to acquiring clinical research knowledge from text has the potential to supplement existing expert-based methods for clinical research standards development and also increase the likelihood of standards adoption.

References

1. Tenenbaum JD, Sansone SA, Haendel M, A sea of standards for omics data: sink or swim? *J Am Med Info*, 2013, in press.
2. Miotto R, Weng C, Unsupervised Mining of Frequent Tags for Efficient Clinical Eligibility Text Indexing, *J Biomed Inform* 2013 Dec;46(6):1145-51.

Creating a Next-Generation Research Informatics Infrastructure: WICER Lessons for Data Integration

Adam Wilcox, PhD[1], Daniel Fort, MPH[2], Suzanne Bakken, PhD, RN[2]
[1]Intermountain Health Care, Salt Lake City, UT; [2]Columbia University, New York, NY

Abstract
We describe lessons learned in creating an informatics infrastructure using patient-centered data, including data from EHRs. These practical and demonstrated lessons will be invaluable for subsequent databases focused on multiple sources for a defined population.

Introduction
Along with increased adoption of electronic health records (EHRs) has been a concurrent interest in using data from EHRs for research. Large semi-coordinated initiatives in the NIH-funded CTSA program, ONC's SHARPn project, NHGRI's eMERGE Network, and AHRQ's PROSPECT studies all have had specific focus on using data from EHRs for research. The Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project at Columbia University is one of AHRQ's PROSPECT initiatives, designed specifically to create an informatics infrastructure to support use of EHR and other data for research. Unique to WICER among all these projects is its comprehensive focus on a population. While the other projects have combined data from multiple sources and at times added other data sources, WICER started with a population and linked data sources specifically relevant to it. WICER includes data from multiple disparate clinical sources, along with patient surveys collected specifically for the project. This patient-centered approach represents a new paradigm of research using EHRs, but introduces some important challenges. In the process of building the WICER infrastructure over the past 3 years, we have navigated around these challenges, adjusting and adapting the system according to the solutions. The WICER project is intended to be a model of this next-generation, patient-centered data infrastructure, so the lessons learned are important as similar projects arise and as existing secondary EHR data use projects become more patient-centered. Our goal is to elucidate the practical discoveries, so that subsequent projects can be more efficiently created.

Methods
Barriers and challenges were encountered in the development of the WICER infrastructure. Issues were considered challenges when either they included complexity that were unanticipated but emerged during the project, or when current approaches were precluded by changes in the environment. To address each challenge, we followed a process to overcome the issue while preserving the overall project goals and integrity. This process was first, to identify and apply established solutions where applicable. Where solutions were not established, we would propose internal solutions. We would then use existing data in the WICER research data warehouse to evaluate the potential impact of potential solutions. Then we would perform pilot studies for a solution, evaluate its effect in the pilot environment, and modify where appropriate. Finally we would apply the solution generally. This standard approach for incremental development was followed consistently for each project.

Results and Discussion
The major challenges we faced and applied solutions in data integration with the project were anonymization, merging, governance issues across sites, centralization, visualization, source population overlaps, cloud security, structured data density, and infrastructure sustainability. We determined that data anonymization was not possible, and approached confidentiality specifically using Safe Harbor rules. Initially we merged identified data and then removed identifiers; our eventual strategy was to distribute a shared research identifier across institutions and then have each institution provide de-identified data using the shared identifier. Similar to many concurrent projects that merged data from different sources, we found data governance issues as the greatest challenge to navigate. Improved methods of de-identification along with federation of data for governance stewardship was needed. Determining source population overlaps determined the relative value of different data sources for specific research questions, which also clarified governance navigation. Data context visualization was necessary for translation to clinical researchers. We also developed a post-collection data modeling approach based on natural densities of structured data. We determined a method for data security using cloud storage, and defined priorities for data sustainability.

Conclusion
Following these lessons can optimize the efficiency of creating an infrastructure to use EHR data for research. Based on our experience, a similar infrastructure could be created with less than half the effort required for our project.

Title: Being PRO ACTive- What can a clinical trials database reveal about ALS?

Neta Zach (PhD, MPA)[1], Robert Kueffner(PhD)[2], Amy Shui(MA)[3], Alexander Sherman(MSc)[4], Jason Walker[4], Ervin Sinani[4], Igor Katsovskiy[4], David Schoenfeld(PhD)[3], Gustavo Stolovitzky(PhD)[5], Raquel Norel (PhD)[5], Nazem Atassi (MD,MSc)[4], James Berry(MD, MSc)[4], Merit Cudkowicz(MD, MSc)[4], Melanie Leitner(PhD)[1]

[1] Prize4Life, Tel Aviv, Israel and Cambridge, MA,USA [2]Ludwig-Maximilians-University, Munich, Germany [3]MGH Biostatistics Center, Massachusetts General Hospital, Boston, MA, USA [4]Neurological Clinical Research Institute, Massachusetts General Hospital, Harvard Medical School, Charlestown, MA, USA [5]IBM T.J. Watson Research Center, Yorktown Heights, New York, USA

Understanding a given patient population is a necessary step in advancing clinical research, improving clinical care for patients, and conducting successful and cost-effective clinical trials. The ability to gather a large enough cohort of patients to achieve that understanding is a challenge, especially in rare diseases such as ALS. The Pooled Resource Open-access ALS Clinical Trials (PRO-ACT) platform was created in order to increase our understanding of the ALS patient population, and its size and scopeprovidethe research community with an unprecedented opportunity.

The PRO-ACT platform consists of over 8600 ALS patients who participated in 17 clinical trials conducted by the industry, non-profit, and government sectors. The creation of the platform required integration of data recorded using different standard of measurements into one comprehensive framework, while maintaining patient anonymity and can save as a valuable case study for other clinical trials platforms. The dataset includes demographic, family history, vital sign, clinical assessment, lab-based , treatment arm, and survival information The database was launched open access to researchers worldwide on December 2012, and since then over 200 researchers from 23 different countries have registered to obtain the data.

Several early assessments were made to start understanding the value of the PRO-ACT database in addressing pivotal questions in ALS clinical research that the large sample size allowed addressing for the first time. These included newly identified predictive features, definitive support for previously proposed predictive features based on small patient samples, and newly identified stratification of patients based on their disease progression profiles. One important initiative included a crowdsourcing effort- the ALS Prediction Prize challenge- to develop improved methods to accurately predict disease progression at the individual patient level. The challenge brought in 1000+ registrants and led to the creation of multiple novel disease progression algorithms tested blindly on a separate dataset. The winning algorithms can aid clinicians and well as substantially reduce the cost of future ALs clinical trials.

These results demonstrate the value of large datasets for developing a better understanding of ALS natural history, prognostic factors and disease variables. in addition the combination of large databases with well controlled large scale crowdsourcing efforts can lead to important novel about ALS natural history, disease progression, patient stratification, disease biomarkers and more. More sophisticated and targeted analysis will reveal additional insights including addressing critical questions about patient stratification and associations with disease co-morbidities and concomitant medications, the identification of biomarkers, and potentially new ways to enhance clinical practice and the design of future clinical trials. Identifying and addressing the most urgent questions that can be answered using this new open-access data resource is of interest.

The Colibri project: a shared database of pediatric patients' examinations

Cristina Altomare[1], Giordano Lanzola[1], PhD, Riccardo Bellazzi[1], PhD, Gianluigi Reni[2]
[1]University of Pavia, Pavia, Italy, [2]IRCCS "E. Medea" – La nostra famiglia, Bosisio Parini, Italy

Abstract
A network of hospitals in Italy launched the Colibri project with the aim to collect magnetic resonance (MR) images and other clinical data of pediatric patients affected by rare diseases (RDs). The goal is combining information from multiple sources to enhance knowledge in the RD domain and speed-up the diagnostic process fostering a collaborative approach.

Introduction
In modern healthcare, collaboration among professionals and information sharing is an important issue [1]. In the context of rare diseases, where existing knowledge needs to be increased, a network supporting communication among geographically distributed actors offers an extraordinary opportunity. Taking advantage of the contribute of the 20 Italian centers of excellence in pediatric neuroradiology, the purpose of the Colibri project is to constitute an indexed archive about RD patients including both cases on which a diagnostic consensus has been reached and cases with an uncertain diagnosis. Symptoms are linked to disorders through a semantic network which is navigated by physicians to verify their diagnostic hypotheses and is cooperatively augmented as they jointly classify complex cases or new disorders.

Materials and Methods
Systems for collecting and integrating multisource and multivariate data are already available [2]. We propose an infrastructure which allows the users to contribute new cases, revise and classify them, exchange opinions or simply consult the database. The physician submits a new case with a tentative diagnosis, after acquiring an MRI scan and visualizing it through the DICOM Viewer. Upon submission, the case, encompassing DICOM images and other clinical data, undergoes anonymization and results in a preliminary recording on both storages (as shown in Figure 1). Then reviewers are notified and may either reject the case or start a review process with the aim of agreeing on a confirmed diagnosis and properly positioning the case into the semantic network according to its attributes. They can also start a discussion using a specific forum made available for each clinical case. Figure 2 shows the complete workflow involved.

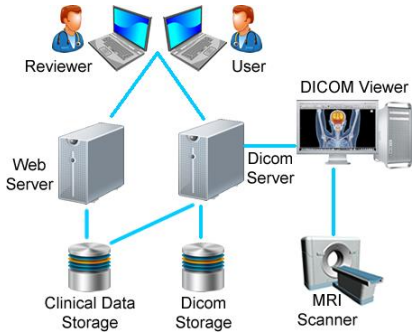


Figure 1: The architecture of the Colibri project.

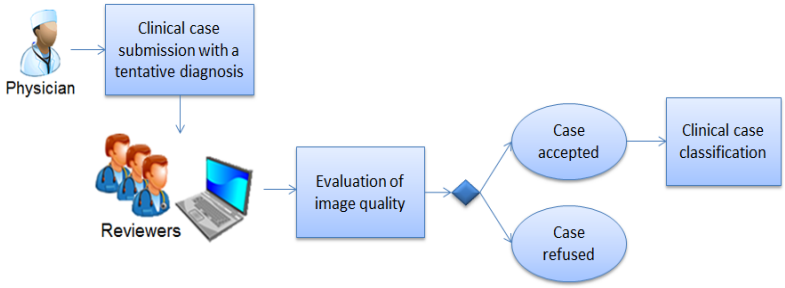


Figure 2: The workflow involved with case management.

Results
We implemented a Virtual Private Network (VPN) among all participants in order to ensure the secure sharing of data. The architecture has been conceived borrowing some open-source tools to avoid their re-implementation from scratch. First we decided to base the DICOM storage on an image archive compliant with the DICOM standard (Dcm4chee). A client-side DICOM Viewer (Mayam), able to communicate with the Dcm4chee for storing DICOM objects, has been customized and endowed with other capabilities, such as data anonymization. A web application, developed using the Google Web Toolkit, supports different functionalities according to the user's role. It allows the consultation of DICOM objects acquired into the DICOM storage and the integration of them with other information such as the case classification at the end of the review process. Finally a web based DICOM viewer (Oviyam), that works with the Dcm4chee, has been integrated into the web application in order to visualize DICOM images on the web.

Conclusion
The paper describes the first Italian infrastructure about RDs in childhood which is presently being tested and will be used by the 20 Italian project partners later this years. Eventually, Colibri will be made available both to experts in RDs to deepen their knowledge and to resident doctors to accelerate the classification of patients with an uncertain diagnosis.

References
[1] J.H. Gennari, C. Weng, J. Benedetti, D.W. McDonald, Asynchronous communication among clinical researchers: A study for systems design, Int. J. Med. Inform. 74 (2005) 797-807
[2] C.G. Fonseca, M. Backhaus, D.A. Bluemke, R.D. Britten, J.D. Chung, B.R. Cowan, I.D. Dinov, J.P. Finn, P.J. Hunter, A.H. Kadish, D.C. Lee, J.A.C. Lima, P. Medrano-Gracia, K. Shivkumar, A. Suinesiaputra, W. Tao, A.A. Young, The Cardiac Atlas Project-an imaging database for computational modeling and statistical atlases of the heart, Bioinformatics, 27 (2011) 2288–229

Automating Data Re-Use Policies for NIH Intramural Clinical Research Data

Elaine J. Ayres, MS, RD ; James J. Cimino, MD,
Laboratory for Informatics Development
NIH Clinical Center, Bethesda, MD

The National Institutes of Health (NIH) comprises 27 individual institutes and centers, many of which conduct clinical research at the Clinical Center, a 240-bed research hospital located on the NIH campus in Bethesda, Maryland. The Clinical Center's electronic health record (EHR) contains myriad data of potential value for secondary use in clinical research. These data are somewhat unique in that they are all collected as part of one or more ongoing interventional or observational studies. A federal mandate requires the lowering of barriers to access to data collected using federal funds.[1] Applying such requirements to Clinical Center data requires balancing the protection of the privacy of human subjects who contributed the data and the intellectual interests of the investigators who collected the data. Data collected in the course of patient care transcend the data sets typically collected in clinical research, so patient records are not analogous to research data sets that are made available through repositories such as dbGAP.

The National Institutes of Health's Biomedical Translational Research Information System (BTRIS) is a repository of intramural clinical research data collected since 1976 in two contiguous NIH EHRs and a variety of clinical trials data management systems.[2] In addition to providing access to identified data from active clinical studies to the investigators on those studies, BTRIS also provides access to data from which identifiers have been removed across all studies. Previously, NIH researchers obtained data from their own and other studies through the EHR or the Medical Records Department. BTRIS provides the opportunity to apply NIH data access policies in a formal, explicit manner. Although the nature of the NIH Clinical Center and its data are somewhat unique with respect to their dual patient care and research nature, many of the policies, and the means for their enforcement, will be relevant to other institutions. For example, many clinical research projects make use of medical center EHRs for order entry and data collection as part of the normal course of patient care. The purpose of this presentation is to describe the how the following NIH policies are implemented in BTRIS:

1) **Patients**: identifiable data are only released to the Principal Investigator (PI) of the study on which the patient is a subject; additional releases of identified data are made at the discretion of the PI to others with documented, NIH-approved roles on the project (associate investigator, closely supervised researcher, data manager). BTRIS users only have access to data collected prior to the end of the subjects' involvement in the study.
2) **Investigators**: Use of data (from which identifiers have been removed) on protocols that are active or terminated less than two years from the time of data access require permission of the original PI prior to use for research purposes. BTRIS users can appeal the PI's decision to a Data Access Committee.
3) **BTRIS users**: Intramural data collected with NIH funds belong to the NIH, which has determined that intramural researchers have the right to immediate reuse for legitimate purposes, subject to the above patient and investigator policies, with oversight by the Office of Human Subjects Research and Protection. BTRIS automates the oversight process.

1. http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
2. Cimino JJ, Ayres EJ, Beri A, et al. Developing a self-service query interface for re-using de-identified electronic health record data. *Stud Health Technol Inform*. 2013;192:632-6.

A Pilot Evaluation of Patient Data Sharing Preferences

Elizabeth Bell, BS, M. Adela Grando, PhD, Lucila Ohno-Machado, MD, PhD
Division of Biomedical Informatics, University of California San Diego, La Jolla, CA

Abstract

We developed a taxonomy of patient preferences for clinical data sharing, and a corresponding graphical user interface for patients to express their choices. Responses from 40 subjects who tested the interface and completed a 28-question survey suggest that our taxonomy contains the necessary elements, and that individuals understand available choices.

Motivation

Many patients may not know how their data are being used for research. Informed consent is typically broad and binary, and tiered approaches to informed consent are seldom utilized. Innovative systems that inform patients about the current use of their data and allow them to select tiered opt-out options may offer advantages in increasing the transparency with which data are shared for research. The purpose of this pilot study was to understand if the educational material and the taxonomies provided in our current prototype were clear and informative enough for participants, and whether an opt-out system for particular clinical data categories (e.g., genetic data) could change their interest in participating in research involving secondary use of their clinical data.

Materials and Methods

A sample of 40 volunteers who responded to recruitment advertisements at the UCSD campus participated in a 45-60 minute session that included use of a graphical user interface (GUI) to select data sharing options, and a 30 minute survey containing 28 questions. Participants logged into the system, which contained 8th grade level educational material on topics such as the types of medical data that can be made available for research, potential risks and potential benefits associated with sharing data, etc. Three sub-taxonomies organized individual choices about data sharing and focused primarily on questions such as "What I am sharing?," "Who am I sharing these data with?," and "What types of institutions do these researchers work for?" An example of the latter is a choice of sharing data with for-profit and non-profit institutions.

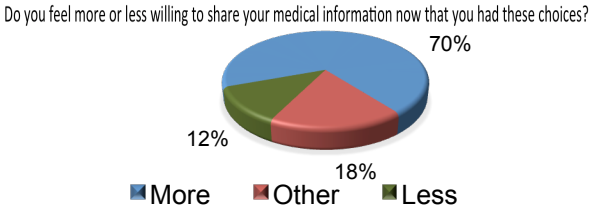


Figure 1. Willingness to share after making choices

I do not want to share:	**Number of participants**
Mental health information	7
Alcohol, substance and smoking information	5
Genetic information	5
Sexual and reproductive health information	5

Table 1. Results on sensitive categories (out of 40 respondents).

Results and Conclusion

Our results suggest that engaging patients in data sharing decisions may be helpful for secondary use of clinical data: 70% of the participants felt that they were more willing to share their medical information after being given choices (Figure 1), and 92% of the participants were interested in knowing more from the researchers who used their data, such as research aims and study outcomes. Additionally, 25% (10) of the participants chose to keep at least one category of sensitive information private (i.e., not shared for research purposes), and most chose more than one category. These category choices are shown in Table 1. Finally, three questions were designed to measure reading comprehension and to check whether the participants thoroughly read the material; the overall score for these questions indicated 90% comprehension.

Acknowledgements: We thank Mona Wong and Claudiu Farcas for technical expertise, and NIH for support from U54HL108460–S2.

Dynamical Approaches to Clinical Artificial Intelligence, Decision-Making, and Cognitive Computing

Casey C. Bennett, MA[1,2], Kris Hauser, PhD[1], Thomas Doub, PhD[2]
[1]Indiana University, Bloomington, IN; [2]Centerstone Research Institute, Nashville, TN

Abstract

We describe ongoing research around temporal modeling of electronic health record data, with specific applications to clinical artificial intelligence (AI) and decision support systems (CDSS). In particular, we focus on recent developments of such approaches in the vein of cognitive computing.

Introduction

Temporal modeling (e.g. sequential decision-making) holds great promise for healthcare, where treatment decisions must be made over time, and where continually re-evaluating ongoing treatment is critical to optimizing clinical care for individual patients. Tremendous advances have been made in data mining and temporal modeling of healthcare data, but practical challenges exist in moving these advances from the laboratory/theoretical setting to applied settings with real patients.

Methods

Previous work has shown the potential for sequential decision-making and reinforcement learning approaches, e.g. Partially Observable Markov Decision Processes (POMDPs) and Dynamic Decision Networks (DDNs), relative to current treatment-as-usual models of healthcare [1]. In the current work, we report on ongoing work to integrate such models into the cognitive domain of clinicians, as well as provide empirical evidence to address a number of practical challenges (e.g. optimal trade-off between treatment costs and outcomes in temporal modeling). In particular, we focus on methods for learning these from clinical data.

Results

First, we provide evidence showing an AI based on sequential-decision making outperforms current treatment-as-usual (TAU) models of healthcare both in terms of cost-effectiveness ($189 vs. $497) as well as outcomes (approx.. 30-35% increase) in a co-occurring chronic illness setting. Second, we provide preliminary results of ongoing research to address a number of practical challenges, including optimal costs/outcomes trade-offs, which may be informative for work in the comparative effectiveness domain. Finally, we discuss new collaborative work between IU, CRI, Regenstrief, Marshfield Clinic, and Wake Forest to address implementation challenges for AI-based CDSS across a range of clinical domains.

Discussion

AI-based tools hold potential to extend the current capabilities of clinicians, to deal with complex problems and ever-expanding information streams that stretch the limits of human ability. Averse to previous generations of AI and expert systems, these approaches are increasingly dynamical and less computationalist – less about "rules" and more about leveraging the dynamic interaction of action and observation over time. The (treatment) choices we make change what we observe (clinically, or otherwise), which changes future choices, which affects future observations, and so forth. As humans (clinicians or otherwise), we leverage this fact every day to act "intelligently" in our environment. To best assist us, our clinical computing tools should approximate the same process. The over-arching goal of the presentation is to stimulate discussion on the topic, and how clinical research informatics can facilitate such development.

References

1. Bennett CC, Hauser K. Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. Artificial Intelligence in Medicine. 2013; 57(1): 9-19.

Methods for Identification of Sexual Health Variables in a Research Data Warehouse

William Brown III, DrPH, MA[1,2], Walter Bockting, PhD[2,3], Nancy Reame, PhD, FAAN[3], Suzanne Bakken, RN, PhD, FAAN, FACMI[1,3]

[1]Department of Biomedical Informatics; Columbia University, New York, NY; [2] Division of Gender, Sexuality, and Health, Columbia University & New York State Psychiatric Institute, New York, NY; [3]School of Nursing, Columbia University, New York, NY

Abstract

The objective of this research was to identify the availability of sexual health-related clinical data through querying the WICER research data warehouse. Of 34 sexual health terms, 5 returned no results and 29 produced 270 clinical concepts of which 230 were unique. Ex: Immunodeficiency TP-2 (HIV-2); Ego-Dystonic Sex Orientation.

Introduction and Background

Little is known about the sexual health of the predominantly Latino population in Washington Heights/Inwood in general, and for those who identify as Lesbian, Gay, Bisexual, and Transgender in particular. Through the Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) project, we are integrating clinical data from NewYork-Presbyterian Hospital (NYP) with survey data from more than 6,000 community residents. Approximately 30% of survey participants identified sexual health, particularly HIV, as one of their top three health concerns. These unique data provide the opportunity to examine the sexual health related needs critical for intervention development. The objective of this research was to identify the availability of sexual health-related clinical data by querying WICER's research data warehouse (RDW).

Methods

First, we established the WICER Sexual Health Working Group (SHWG). Its eleven members contributed query terms from their specific sexual health discipline via shared Google Docs spreadsheets, and entries were categorized as Laboratory, Diagnosis, Procedure, Medication, Location, or Clinical variable. Then, we leveraged four major functions of Research Data Explorer (RedX) to identify sexual health-related clinical variables in the RDW: 1) query RDW, 2) save query results, 3) save query parameters, and 4) develop a library of key variables. No actual patient or research participant data was searched. The query results from RedX - including Medical Entities Dictionary (MED) codes and ICD-9 CM codes - were then added to the Google Docs spreadsheet. We used pivot tables to aggregate similar and synonymous variables by MED and ICD-9 CM codes. What's more, the tables allow for bidirectional traversing of the MED to identify other relevant clinical concepts.

TERMS	CLINICAL CONCEPTS	ICD9	MED
Contraception	HX OF CONTRACEPTION	V15.7	22375
Gender	GEND IDEN DIS,ADOL/ADULT	302.85	7695
Gonorrhea	ACUTE GC INFECT LOWER GU	98	6929
HPV	ABN PAP CERVIX HPV NEC	795.09	30007
Immunodeficiency	HIV COUNSELING	V65.44	41659
Immunodeficiency	IMMUNODEFICIENCY TP 2 (HIV 2)	79.53	40140
Menopause	PREMATURE MENOPAUSE	256.31	30007
Reproductive	ASSIST REPRO FERTILITY	V26.81	30007
Reproductive	PREG W POOR REPRODUCT HX	V23.5	22522
Sexuality	EGO-DYSTONIC SEX ORIENT	302	7666

Table 1. Examples of input terms, and clinical concept variables/codes generated by RedX

Results and Discussion

The SHWG members contributed 34 key sexual health terms. Five terms returned no results in RedX (e.g. Kallmann Syndrome). However, alternative general terms such as "Puberty" produced synonymous Kallmann Syndrome clinical concepts (e.g. Delay Sexual Develop NEC). The remaining 29 terms produced 270 concepts of which 230 were unique (For examples see Table 1). We plan to expand concept retrieval through traversing the MED hierarchy. Findings from the RedX/RDW clinical concept discovery process will be used to query the clinical data warehouse, and will also form the basis for the development of future studies to advance scientific knowledge and promote the sexual health of minority populations including those who identify as Lesbian, Gay, Bisexual, and Transgender.

Acknowledgments: R01HS019853, R01HS022961, NYS Department of Economic Development NYSTAR (C090157). Dr. William Brown III is supported by 5T15LM007079-22.

Supporting the Discoverability of Research Objects by Connecting Research and Researchers with ORCID

Rebecca Bryant, PhD, ORCID, Bethesda, MD and Kristi L. Holmes, PhD, Washington University School of Medicine, St. Louis, MO

Abstract

A long-standing challenge within the research community has been the inability to reliably connect individuals with their contributions, including articles, datasets, and other research objects. Researchers also struggle to make connections with potential collaborators. ORCID[1], an open, non-profit, and community-driven organization, provides a unique and persistent identifier to researchers, connecting them with their activities through integration in research workflows. Since its launch in October 2012, the ORCID registry has grown steadily and organizations within the interconnected publishing, funding, and academic research communities have integrated the ORCID identifier into their workflows. Support from the Alfred P. Sloan Foundation has enabled ORCID to support the integration with variety of research platforms, including VIVO, DSpace and Hydra/Fedora repository tools, research data life cycle management tools like HubZero, and the Reactome biological pathways knowledge base data center.[2] This poster will provide an overview of ORCID and how ORCID iDs can be obtained and profiles curated. The poster will also provide examples of how platforms like VIVO[3], Reactome[4], and HubZero[5] are integrating ORCID iDs into their systems, better enabling the open collection, dissemination, archiving, and discovery of a broader range of research objects as well discovery and visibility of the researchers who create them.

[1] http://orcid.org/
[2] http://orcid.org/blog/2013/09/27/announcing-orcid-adoption-integration-program-awardees
[3] http://vivoweb.org/
[4] http://www.reactome.org/static_wordpress/about/
[5] http://hubzero.org/

Evaluating Clinical Studies for Scientific Merit and Financial Feasibility Prior to Human Subjects Review: A Workflow Approach Using REDCap

Thomas R. Campion, Jr., PhD, Daniel Izcovich, BA, Vanessa L.I. Blau, BA, Scott W. Brown, MA, Jaclyn E. Fronda, BS, Scott N. Robertson, BA, Aleta R. Gunsul, MPA, Erica E. Love, MA, Alicia N. Lewis, MA
Weill Cornell Medical College, New York, NY

To assure high quality research, institutions can review scientific merit and financial feasibility of clinical studies in addition to human subject protection plans. To facilitate such a process, we rapidly developed an approach using REDCap and custom scripts embedded in existing workflow. Preliminary results indicate overall user satisfaction.

Introduction

Increasing the volume and quality of clinical research while maximizing utility of limited resources is a goal of academic medical centers. In January 2013, Weill Cornell Medical College of Cornell University (WCMC) and NewYork-Presbyterian Hospital (NYPH) established a Joint Clinical Trials Office (JCTO) to grow the clinical research enterprise of both institutions. As of August 15, 2013, the JCTO began requiring WCMC investigators to submit all clinical studies for scientific merit and financial feasibility review by the Clinical Study Evaluation Committee (CSEC) prior to human subjects review by the Institutional Review Board (IRB). To support the CSEC submission process while enabling completion of IRB applications in parallel, JCTO on July 18, 2013 requested the WCMC Research Administration Computing (RAC) unit rapidly implement an automated electronic solution.

Methods

JCTO and RAC developed an approach embedded in researcher workflow using REDCap 5.6.4 and custom scripts. First, we configured the electronic IRB system (eIRB) to only allow CSEC administrators to create IRB protocols. Second, we redirected hyperlinks for investigators creating new protocols in eIRB to a new CSEC REDCap survey. Third, we created a REDCap project consisting of three data collection instruments: "Part A," a survey for investigators submitting basic characteristics of a study (e.g. title, review type, investigator name) accessible via the aforementioned links; "Part B," a form restricted to the investigator specified in Part A for a particular study to provide study details; and "CSEC Administration," a form restricted to CSEC administrators to manage study approval following submission of Parts A and B. Fourth, we created scripts using PHP 5.3 and the REDCap API that, upon investigator completion of a study's Part A or Part B, emailed CSEC administrators validated REDCap submission data in a format tailored to completing next steps. For Part A, next steps included verifying user input, copying-and-pasting basic study characteristics from REDCap into eIRB to create a new protocol, pasting the IRB protocol number from eIRB into REDCap, assigning the investigator to a REDCap data access group specific to the investigator of a submitted study, and emailing the investigator a link to Part B as well as the IRB protocol number so the investigator could begin completing Part B and the eIRB application in parallel. For Part B, next steps included verifying user input and preparing REDCap-generated PDFs for weekly in-person CSEC review meetings. After review meetings, CSEC administrators recorded committee decisions on each study's CSEC Administration form. For approved studies, custom scripts retrieved IRB protocol number, funding source, and short title from REDCap and transferred the data to the clinical research management system, which previously received basic study characteristics through an interface with eIRB. As an alternative to REDCap submission, investigators could also submit a PDF version of Parts A and B that CSEC Administrators transcribed to REDCap.

Results and Discussion

JCTO and RAC deployed the solution on August 15, 2013. Through October 1, 2013, investigators have completed 150 Part A and 51 Part B submissions. CSEC has approved 25 studies overall, scheduled 20 for review, and required revision for 6. Investigators have submitted all but one study via REDCap and reported being satisfied overall. CSEC administrators have reported overall satisfaction, especially using REDCap to generate ad hoc reports, but that copying-and-pasting between systems is time consuming. REDCap has proven to be an effective platform for rapidly supporting a new review process with minimal disruption of existing workflow. Future work will evaluate effects of CSEC review on volume, quality, and other measures of the clinical research enterprise.

This work received support from UL1 TR000457-06 awarded to WCMC Clinical and Translational Science Center.

Open Source Integrations: How Fred Hutchinson Cancer Research Center connected their Drupal Resource Collection to the eagle-i Network

Ann Marie Clark, MLS[a], Bhanu Bahl, PhD[b], Daniela Bourges-Waldegg, PhD[b], Aaron Lamb, MFA[c], Beth Levine, BA[a], John Locke, BA[c], Julie McMurry, MPH[b], Ann Reynolds, PhD[a], David Tolmie, MLIS[a] and Douglas MacFadden, MS[b]

[a]*Fred Hutchinson Cancer Research Center, Seattle, WA.* [b]*Harvard Medical School, Boston, MA.* [c]*Freelock, Seattle, WA.*

Abstract

Recently, the Fred Hutchinson Cancer Research Center decided to share the contents of its Shared Resources website with the eagle-i Network. A joint FHCRC/Harvard/Freelock team successfully developed a reusable open source solution to connect the two systems.

Introduction and Background

eagle-i (www.eagle-i.net) is a biomedical resource sharing network that currently counts 26 institutions sharing more than 54000 resources. The eagle-i platform gives investigators the ability to easily discover resources that can enhance and accelerate their research. Developing or sourcing such resources in each individual laboratory would otherwise be inefficient, costly, or in many cases impossible. In 2012, the Arnold Library at the Fred Hutchinson Cancer Research Center in Seattle, WA contacted the eagle-i team to explore technical options for joining the network. This world-class cancer research center recognized the benefits of visibility on the eagle-i Network, but faced the challenge of connecting their existing Drupal-based resource information to eagle-i without introducing inefficiency or duplication of labor.

Methods

It was clear that the only way to manage the cost and efforts involved in maintaining content in both the Shared Resources website and eagle-i would be to connect the two systems in an automated fashion. The team (FHCRC, Harvard and Freelock) devised a solution that leverages eagle-i's distributed architecture: an FHCRC eagle-i node was installed locally and through a custom Drupal module, automatically populated with content from the Shared Resources website; after this, joining the eagle-i network was transparent, as it simply required eagle-i's central search application to index the FHCRC eagle-i node in the same manner it indexes all other nodes. The team developed the custom Drupal module to handle the data flow from FHCRC to the eagle-i local node. Designed to reside alongside the existing Shared Resources website, the new module maps the site's content to the eagle-i ontology and creates resource descriptions in the local eagle-i node. The team also developed a new eagle-i web service that accepts resource descriptions generated by third party applications, and that is contacted by the Drupal module when new content is created or when existing content is updated. The system allows the content to be created or edited in only one location (the Drupal site) and once published, key elements of that same material are instantly available in the eagle-i network.

Results and Discussion

This open source solution allows seamless propagation of descriptions and metadata from the Fred Hutchinson Shared Resources website to the eagle-i Network without duplication of data entry or manual data integration. The resulting Drupal module code base has been shared as open source in the Drupal.org community site, and the new web service is now part of the standard eagle-i open source software stack. This Drupal module can be a useful option for other Core Labs and Shared Resources groups that are looking to partner with the eagle-i network. Future enhancement opportunities include the development of a schema-mapping user interface, as the module currently relies on a static XML map that translates the Drupal site data model to the eagle-i ontology. A more detailed technical description of the project is available on Freelock's website. (Lamb 2013)

Acknowledgements/Funding

FHCRC and Freelock were supported by Federal funds to Ann Marie Clark from the DHHS, NIH, NLM, under Contract No. HHS-N-276-2011-00008-C with the University of Washington. eagle-i was originally supported by an ARRA award from NCRR, NIH to Dr. Lee Nadler (#U24 RR 029825).

References

1. Lamb, Aaron. *Drupal and the Semantic Web - Introducing the eagle-i Drupal module.* July 29, 2013. http://www.freelock.com/blog/aaron-lamb/2013-07/drupal-and-semantic-web-introducing-eagle-i-drupal-module (accessed October 2, 2013).

Bridging Clinical and Research Data through Informatics: Combining Automated Clinical Data Abstraction with Manual Annotation

Bas de Veer, MS[1], Christine Fong, MS[1], Sally Lee, PhD[1], Chris Nefcy[1], Sarah Prager, MD[2], Tony Black, MA[1]

[1]University of Washington, Institute of Translational Health Sciences, Seattle, WA

[2]University of Washington, Department of Ob/Gyn, Seattle, WA

Abstract

We have developed a bridge for securely moving data between a clinical data repository and a research electronic data capture system (REDCap). Use of the bridge has greatly reduced the time and resources required to collect and organize the data needed for research studies.

Introduction and Background

Informatics solutions to address the needs within the clinical research community have led to the creation and use of several classes of clinical systems such as clinical data depositories, for integration of patient data across multiple sources and electronic data capture systems, for collection and management of research data. Although these systems are typically utilized separately, building a bridge between empowers researchers to automate patient data collection into study databases that can be easily managed and manually annotated. We have developed a bridge between an electronic data capture system (REDCap) and a clinical data repository which are both widely used at the University of Washington.

Methods

The University of Washington Clinical Data Repository (UWCDR) resides in a secure server space that is disconnected from the local REDCap database. We created a custom C# library (REDCap-CDR bridge) that can access data securely from both REDCap and an external repository through the REDCap API while using SQL Server Integration Services (SSIS).

Results

We have successfully utilized the REDCap-CDR bridge for a study of the effects of Long Acting Reversible Contraception (LARC) in various patient populations. The bridge was used to import patient IDs entered into a REDCap project into the UWCDR, query necessary patient data, and then export the results back into REDCap for manual annotation (Figure 1). Compared to manual abstraction, automatic extraction of data from the clinical data repository saved time and created a partial dataset that could be analyzed relatively early in the research study. The reduction of manual abstraction also reduced transcription errors.

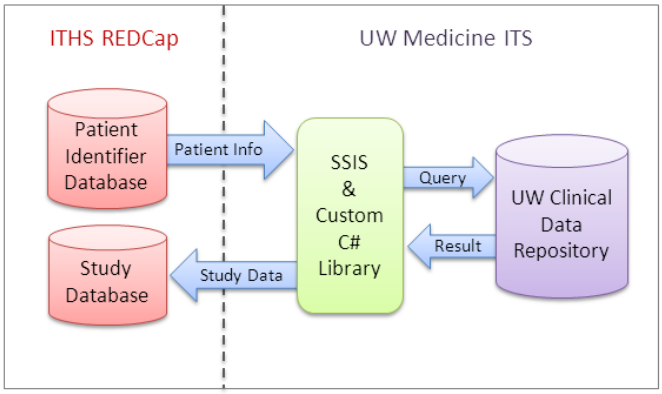


Figure 1. Dataflow within the LARC study utilizing the REDCap-CDR bridge.

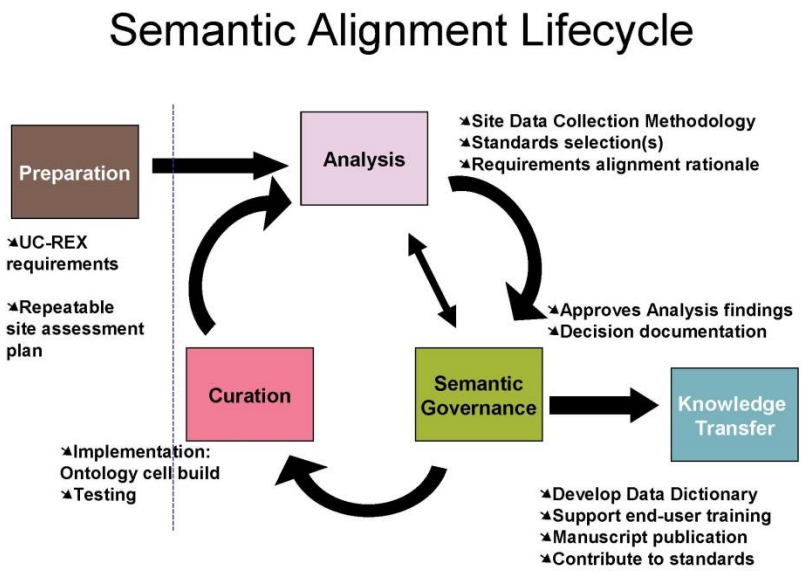
Discussion

In our studies, the REDCap-CDR bridge has helped to connect two separate clinical research workflows into one, saving time and effort. Our solution has limitations in that unstructured clinical data (e.g. free text, clinical notes) and incomplete UWCDR data still necessitated some manual abstraction and validation. However, partially automated data extraction was still preferable to a complete manual abstraction.

UCReX Data Harmonization: A Report from Three Years of Data Alignment Supporting Research Across Five University of California Health Systems

Davera Gabriel, RN[4], Dana Ludwig, MD[5], Douglas Bell, MD PhD[1], Paulina Paul, PhD[2], Whenhong Zhu, PhD[2], Ayan Patel, MS[1], Travis Nagler, MS[4], Douglas Berman, Lisa Dahm, MS[3]
[1]University of California, Los Angeles, [2]University of California, San Diego, [3]University of California, Irvine, [4]University of California, Davis, [5]University of California, San Francisco

The University of California Research Exchange (UCReX) is a program aimed at enhancing collaboration among biomedical researchers at the five UC health system campuses - Davis, Irvine, Los Angeles, San Diego and San Francisco. Albeit the overarching program is technical platform agnostic, UCReX is currently utilizing the i2b2 framework to support a federated implementation exposing characteristics of the aggregate UC service population of nearly 13 million Californians. Central to a federated i2b2 implementation is aligning data from heterogeneous sources to a harmonized standard in order to support execution of a single query against all participating nodes. The UCReX Data Harmonization working group, with representatives from each of the 5 campuses, utilizes the Semantic Alignment Lifecycle approach to developing the target ontology which acts as a semantic "hub" which permits semantically aligned queries to access source data from all feeder systems to populate the federated i2b2 environment Our submission aims to discuss the results of our data harmonization work, outlining the pitfalls, compromises and best practices in achieving semantic alignment.



Issues addressed include implementation of the United States Federal Office of Management and Budget (OMB) 1997 reissued set of standards for the classification of federal data collected on race and ethnicity, as well as issues associated with other demographic elements (gender, marital and vital status…) which provided challenges for our research secondary use scenarios, requiring tabulation of population demographics data from divergent health care and research sources. Similar alignment and tabulation issues have been encountered in the selection of data standards that comprise the semantic target or "hub," such as utilizing the ICD-9-CM administrative classification to represent clinical diagnoses and procedures data; plans for the advent of ICD-10, normalization of laboratory results and the utility of LOINC as an alignment coding system, and issues associated with representing medications.

One of the pervasive experiences the working group has had thus far is that there are numerous implications for supporting the research enterprise as a whole that need to be considered when selecting an approach to harmonizing data created in heterogeneous sources and settings. Any or all of these considerations may have a role in determining the best fit in attaining semantic alignment appropriate for project functional requirements, technical platform, available (personnel) resources, as well as an evolving research policy milieu. Additionally, the group has encountered success by balancing the anticipated needs of clinical research end-users by defining the scope of harmonization work to be accomplished in annual cycles and prioritization of data that are high-value to a large group of clinician researchers and / or highly available. These approaches, combined with utilizing the Semantic Alignment Life Cycle have created a project product that accomplishes some, if not all, of the program objectives with available resources and time constraints.

DELVE: A Document Exploration and Visualization Engine

Daniel R. Harris[1], Ramakanth Kavuluru[2], Stanley Yu[2], Robert H. Theakston[2], Jerzy W. Jaromczyk[1], Todd R. Johnson[2]

[1]Department of Computer Science, College of Engineering, University of Kentucky, Lexington, KY 40506.
[2]Division of Biomedical Informatics, Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, KY 40506.

Brief Summary: We present DELVE (Document ExpLoration and Visualization Engine), a prototype for performing literature-based searches with the aid of interactive visualizations and a framework for quickly implementing such visualizations as modular Web-applications.

Introduction and Background: The goal for DELVE is to better satisfy the information needs of researchers and help them explore and understand the state of research in scientific literature. For proof of concept, we implemented DELVE using two years of PubMed citations made available by the NLM. Currently, many search engines, including PubMed, return a linear list of search results. A DELVE search returns instead an interactive set of panes through which different visualizations enable insightful data exploration; researchers delve into their query so that relevant results more quickly enter into their vision and irrelevant results can be systematically removed and dismissed.

Methods: In our work, we have generalized the concept of a word cloud, a visualization rendering words with size relative to frequency, to also include renderings of phrases and MeSH terms. Additionally, we provide a word tree to show context and a documents list that can be refined interactively. We base our prototype on principles from user-centered design and human-computer interaction (HCI), in particular Shneiderman's information visualization mantra: overview first, filter and zoom, provide details on demand. The interaction in DELVE is along two different axes: (1) adaptive linking between different visualization widgets within its framework and (2) linguistic alignment features allowing user and tool to collaborate on lexical and semantic views of information sources, rather than forcing the user to adapt to the tool--a technique we call cooperative semantic information processing.

Results and Conclusions: We believe these principles are instrumental in developing a literature-based search tool that is capable of addressing the complex information needs of modern researchers. Preliminary evaluations demonstrate the usefulness of DELVE's techniques: (1) a clinical researcher immediately saw that her original query was inappropriate simply due to the frequencies displayed in the clouds and (2) a muscle biologist quickly learned of vocabulary differences found between two disciplines that were referencing the same idea, which we feel is critical for interdisciplinary work.

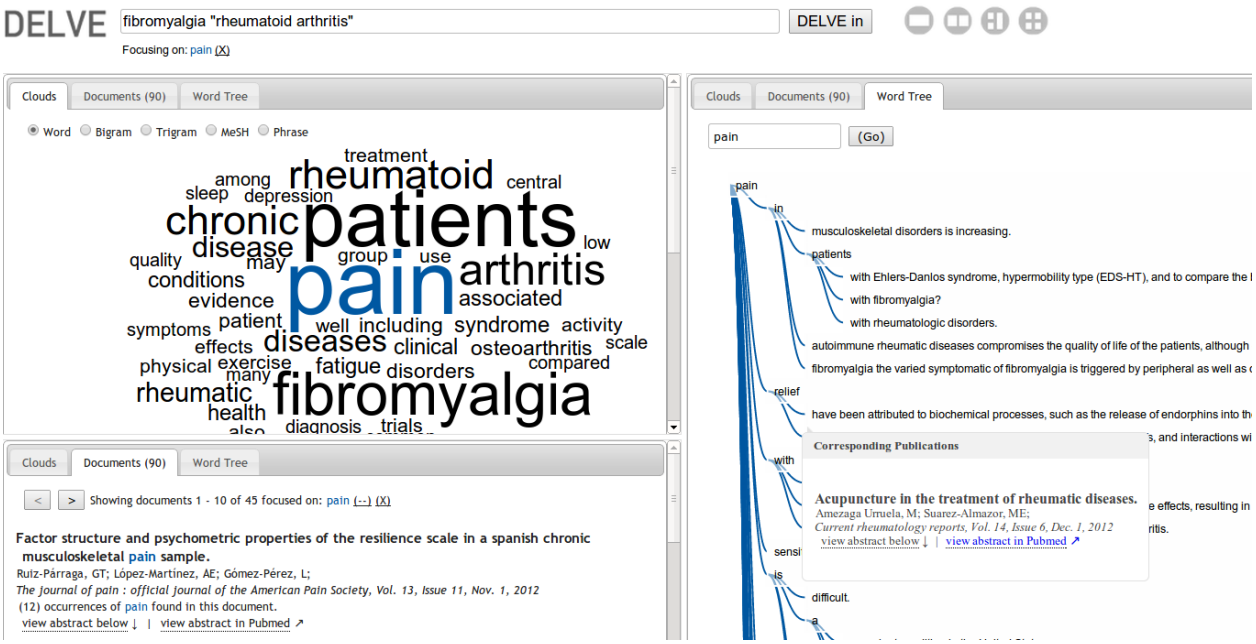


Fig 1. DELVE in action

Comparative Analysis of Online Health Information Search by Device Type

Ashutosh Jadhav, MS[*1], Jyotishman Pathak, PhD[2]
[1]Wright State University, Dayton, OH; [2]Mayo Clinic, Rochester, MN

Abstract

To study online health information search behavior from smart devices (smartphones, tablets) and personal computers (desktop, laptop), we performed a comparative analysis of large-scale health search queries from Web search engines to Mayo Clinic's consumer health information website.

Introduction and Background

Since last decade, percentage of people using Internet to search and learn from the health-related information is increasing exponentially. According to Online Health 2013 Pew Survey, one in three American adults searched online to get information about a medical condition. With recent exponential increment in smart devices ('SD': smartphones and tablets) usage, percentage of people using smart devices for health information search is also growing rapidly. User experience for online information search varies with device used for search such as smart devices and personal computers ('PC': desktop, laptop). Understanding the effect of device used (SD vs. PC) for health information search would help us to learn more insights about health search behavior. Such knowledge can be applied to improve the search experience, as well as develop more advanced next-generation knowledge and content delivery systems.

Methods

Based on the number of visits and the type of device used (PC or SD), we have collected top one million health search queries between June 2011 – May 2013 that direct Online Health Information Seekers (OHIS) to Mayo Clinic.com web pages. MayoClinic.com is one of the top online health information providers and highly ranked (often in top 3) in online health/medical information search. We performed the following analysis on this data: 1) Identify top search queries 2) Categorization of the search queries into health categories such as symptoms, causes, treatment, diet, etc. 3) Average number of words, characters used in the search queries and their range distribution 4) Usage of query operators (such as 'and', 'or', etc.) and special characters in the search queries 5) Expression of information need (using keywords, Wh-questions, Yes/No questions) while formulating search queries, and 6) Misspellings in the search queries.

Results and Discussion

Our analysis leads to the following observations: Google is a leader in online search and our analysis confirms Google's dominance in health information search. Health information searched via different device differs as much as 65%. In one year, health information searched changes by 50% and the change is even higher considering device type. Symptoms, causes and treatments & drugs are top searched health categories respectively. Health search queries are longer than general search queries, which imply that OHIS describes health information need in more detail. Interestingly, health search queries from SD are longer than that from PC. Usage of special characters is limited and it is more for health search queries from PC than from SD. Use of query operator in health queries is less and variation of AND (AND, &, +) is used more often followed by OR and '+'. Operator usage is slightly higher in health queries from SD as compared to that from PC. OHIS formulate search queries primarily using keywords followed by Wh-Questions and Yes/No Questions. OHIS ask more health questions from SD than PC. In Wh-questions, OHIS mostly use What and How in search queries and both of them generally signify more descriptive information need. OHIS ask more temporal questions (When) form SD than PC. OHIS generally use Yes/No questions to check some factual information and most of them start search queries with Can, Does and Is. OHIS ask more Yes/No questions starting with 'Can' using SD than PC. Approximately 1 in every 4 queries has at least one spelling mistake and spelling mistakes in health queries are slightly higher from SD than that from PC.

Conclusion

We observed that health information search behavior differs with device used for search (smart device vs. personal computer.) This study extends our knowledge about online health information search behavior and provides interesting and valuable insights useful for Web search engines, health websites and health providers; and eventually to empower OHIS.

---

[*] This work was done during author's internship at Mayo Clinic, Rochester, MN, United States.

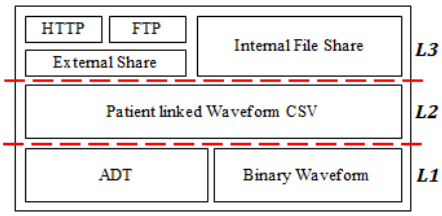
ICU-OR Waveform Data Collection and Repository

Hyeon Joo, Henry Lee, Peter Bow, James Blum, MD
University of Michigan Health System, Ann Arbor, MI

Abstract
We developed the University of Michigan Waveform (MiWave) to capture and store the waveform data on a 24/7 basis for intensive care units and operating rooms. Three main components of the MiWave development are (a) data collection from patient monitors, (b) waveform file linked to patient and (c) file sharing and distribution. Providing patient waveform data of entire stays and transition of location information, researchers are able to analyze symptoms of waveform collected from different bed sites with different care and operations.

Introduction
The needs of data in computable format that researchers can apply sophisticated signal processing, machine learning or artificial intelligent techniques to improve quality and cost of patient care have tremendously increased. The traditional way of storing waveform data, however, is in scanned images, PDF files, or proprietary format for human access to the data. Two major challenges involved in archiving waveform data in computable format for all monitored locations at the University of Michigan are 1) an efficient way of collecting and storing lossless waveform data and 2) linking different locations of patients such as moving to and from an ICU bed to an OR waveform data collected from different monitors.

Methods
To design and develop the *MiWave* repository, we selected UMHS cardiovascular center (CVC) ICU 24 beds and 10 ORs as our first pilot to continuously capture waveform data on a 24/7 basis. The process of *MiWave* at the CVC's 34 beds has three layers: *waveform data collection, patient linked waveform file generation, and waveform file sharing and distribution*.



The *waveform data collection* layer captures and stores binary waveform data from patient monitors, and patient location from Admit, Discharge and Transfer (ADT) messages of the electronic health record (EHR) systems. Captured data consists of several different waveform types: ECG sampled at 240Hz, Art Line, Pulmonary Artery (PA) and Central Venous Pressure (CVP) sampled at 120Hz and Respiratory, SpO2 and EtCO2 sampled at 60Hz. Data points are transmitted in two byte lengths, providing computational resolution. Additional information such as timestamps, sequence id, and the number of data points is added to preserve data accuracy and help patient linking process.

The *patient linked waveform file generation* layer generates the waveform files in CSV format that can be easily imported to analytical tools such as Matlab, SPSS, SAS, EXCEL, etc. A patient census is created in this layer to link the binary waveform data to patients, and then stored into a relational database. Key data elements such as patient location (unit, room and bed) and bed in and out times are cross-referenced and matched to the collected waveform data and formatted into final CSV files. These files are identified by patient, waveform type and capture period.

Lastly, *waveform file sharing and distribution* layer facilitates the waveform file access and delivery to the research community. At this time, various facilities are used and include FTP and direct network access. A full featured waveform file distribution web site is being developed.

Results
MiWave has been successfully in operation for over 10 months at the UMHS CVC 24 ICU beds and 10 ORs. Daily average disk utilization of binary data over the 10 different waveform types is 1.56GB/day for 24 ICU beds and 0.64GB/day for 10 operation rooms. For 24/7 services, extrapolating to entire scale at the University of Michigan would be 4.5TB/year for 179 ICU beds and 2.0 TB/year for 82 operating rooms.

Discussion
Future enhancements for the *MiWave* include features to load key indicators from EHR to help waveform analysis, efficient disk storage utilization with compression, searchable features based on demographics, lab results, etc., and scalable storage model with clinical data from EHR systems.

Stanford-NIH Pain Registry: Open source platform for large-scale longitudinal assessment of clinical data and patient-reported outcomes

Ming-Chih Kao, PhD, MD, Stanford Hospital and Clinics, Karon Cook, PhD, Northwestern University, Garrick Olson, BS, Teresa Pacht, BS, Beth Darnall, PhD, Susan C. Weber, PhD Sean Mackey, MD, PhD, Stanford Hospital and Clinics and Stanford University

No disclosures, Funding NIH HHSN 271201200728P and Redlich Pain Endowment

Introduction: The Institute of Medicine (IOM) in Relieving Pain in America report (2011) called for the development of national patient registries to support the development of learning healthcare systems. In particular for the management of patients with chronic pain, the IOM has called for national patient outcome registries that can support point-of-care decision making and large-scale assessment of safety and effectiveness of therapies.

Methods: Web-based applications are developed to assess patients and to support staff with integrating the Registry into clinic workflow. Patient assessment can be done at home via emailed link, or in-clinic using tablets prior to an appointment. Patient assessments are completed via a web browser, implemented using jQuery Mobile inside Google Web Toolkit, with questions dynamically generated by server. Assessments include multiple types of questions such as numeric scales, selecting areas on body map images, radio lists, checkbox lists, conditional questions that reveal additional questions, among others. Computerized adaptive testing engine is available via Northwestern's PROMIS API and SNAPL-CAT. The clinic interface is implemented using Google Web Toolkit, with client and server written in Java. Appointment schedules are loaded daily from the results of an EPIC report. At patient check-in, in-clinic assessments on iOS and Android tablets may be initiated as needed, using the respective built-in mobile web browsers. All patient surveys and other data are stored in Oracle databases and accessed via SQL over standard JDBC interfaces.

Results: Since roll-out in August 2012 and the subsequent slow ramp-up, over 2,200 unique patients have completed surveys, with over 4,500 assessments overall.

Conclusions: An open source, extensible platform was created that enables rapid definition and deployment of data capture tools. This represents a successful partnership between the NIH and Stanford with funding from most of the NIH Institute Directors.

Automatic Clinical Note Type Classification for Heart Failure Patients

Youngjun Kim[1,3], Jennifer Garvin[2,3], Julia Heavirland[3], Stéphane M. Meystre[2,3]
[1]School of Computing, [2]Department of Biomedical Informatics, University of Utah,
[3]VA Health Care System, Salt Lake City, Utah

Abstract: *The type of clinical notes is important information when assessing the context of clinical information extracted. The type information is sometimes not mentioned in or in association with the note. In these cases, automatically determining the type of clinical note would help to select relevant notes and to analyze medical concepts for patient care. Our machine learning-based classifier achieved 92.23% F_1-measure with 10-fold cross-validation when automatically classifying clinical notes to nine different types.*

Introduction: Clinical note type is important information when assessing the context of clinical information and measurements. For example, for patients suffering from heart failure (HF), diagnostic measurements such as the left ventricular ejection fraction is considered more accurate in echocardiogram reports or cardiology consultation notes than in other types of clinical notes. To determine if a patient is taking certain medications at the time of discharge, the discharge summary is the most accurate source of information. Note type is sometimes not mentioned in or with the clinical note, requiring applications of Natural Language Processing to automatically classify notes for information analysis or specific clinical notes retrieval.[1-2] For this study, we used supervised machine learning to classify clinical notes into nine different categories based on their content.

Methods: We created a random sample of a variety of inpatient clinical notes from patients with HF treated in 8 different VA medical centers in 2008. All 5,059 clinical notes in this sample were categorized into nine different document types: cardiology consultation (card), discharge summary (disc), echocardiogram report (echo), history and physical examination (h&p), nursing note (nurs), other consult (o_con), pharmacy medical reconciliation (medi), pharmacy other (phar), and progress note (prog). For the automatic classification, we implemented a multi-class classifier based on Support Vector Machines (SVM) to determine the type of each note. The classifier used feature vectors composed of a bag-of-words representation of the note text.

Results: Recall (R), precision (P), and the F_1-measure (F) were measured when comparing the automatic note type classification with a reference standard. These metrics are displayed in the table as a confusion matrix. Our classifier reached an overall 92.23% F_1-measure with ten-fold cross validation. Progress notes were misclassified more than other note types, often as cardiology consultation notes with 58 false positives and 64 false negatives.

	Classified as									**Notes**			
Note type	card	disc	echo	h&p	nurs	o_con	medi	phar	prog	**count**	**R**	**P**	**F**
card	538	4	2	23	5	8	1	1	64	646	83.28	80.90	**82.07**
disc	8	627	0	2	1	0	0	0	9	647	96.91	98.12	**97.51**
echo	6	1	100	0	0	0	0	0	5	112	89.29	93.46	**91.32**
h&p	27	1	2	473	1	4	0	1	26	535	88.41	89.08	**88.74**
nurs	14	1	0	2	857	0	0	0	6	880	97.39	98.96	**98.17**
o_con	13	1	1	8	1	192	0	7	25	248	77.42	88.48	**82.58**
medi	0	0	0	0	0	1	70	3	1	75	93.33	95.89	**94.59**
phar	1	0	0	0	0	1	2	528	0	532	99.25	97.06	**98.14**
prog	58	4	2	23	1	11	0	4	1281	1384	92.56	90.40	**91.47**
All	665	639	107	531	866	217	73	544	1417	5059	92.23	92.23	**92.23**

Conclusion: Our work shows that clinical note types can be successfully classified automatically with simple lexical features based on the note textual content, even among heterogeneous note types. This automated system can help filter out irrelevant documents and support more accurate patient level classification.

Acknowledgments: Research supported by VA HSR&D IBE 09-069 (ADAHF) and by HSR&D HIR 08-374 (Consortium for Healthcare Informatics Research) and HIR 09-007 (Translational Use Case – Ejection Fraction).

References

1. Patterson O, Hurdle J. Document clustering of clinical narratives: a systematic study of clinical sublanguages. AMIA Annu Symp 2011:1099–107.
2. Shiner B, D'Avolio LW, Nguyen TM, Zayed MH, Watts BV, Fiore L. Automated classification of psychotherapy note text: implications for quality assessment in PTSD care. J Eval Clin Pract. 2012;18:698–701.

HL7 SS-MIX standard storage using MongoDB

Eizen Kimura, B.M, Ph.D[1], Ken Ishihara, MD, PhD[1]
[1]Dept. Medical Informatics Medical School of Ehime University, Ehime, Japan

Abstract

The standardized structured medical information exchange (SS-MIX) uses the Health Level 7 (HL7) standards to manage the repository, but lacks scalability because it uses a simple file system. We built a virtual SS-MIX storage system using MongoDB, and simulated the file system using the filesystem in userspace (FUSE) module. The performance of a single node was sufficient for university hospital requirements, and showed potential for SS-MIX storage on multiple nodes with the necessary scalability for hosting a nationwide repository.

Introduction

The accumulation of laboratory test results and drug prescriptions throughout Japan represent data that may contribute to clinical research. The standardized structured medical record information exchange (SS-MIX)(1) is an emerging standard for health information exchange in Japan. SS-MIX is an archiving method, allowing storage of Health Level 7 (HL7) messages in an organized file system structure. Physicians can exchange health information by placing HL7 messages in removable storage; however, the simplicity of the SS-MIX standard will be lost in the scaling that is required to create a national repository because it uses a flat file system. We have developed a virtual file system to enable an SS-MIX repository with large-scale data accumulation and high-speed searching.

Method

We developed a tool that maps the metadata from SS-MIX (including the patient ID, directory structure, and HL7 message types) and the HL7 2.x messages into a single binary Javascript object notation (BSON) message, which is stored in MongoDB. We also developed a filesystem in userspace (FUSE)(2) module, which mounts the emulated SS-MIX storage using the contents of the MongoDB database. The performance was evaluated using a system with an Intel Core i7 2.7GHz processor, 16 GB of RAM, a 751 GB SM768E solid-state drive, running MacOS X 10.8.5 with MongoDB 2.4.6. Ten million HL7 messages were generated, which contained randomly generated laboratory test results, and were inserted into MongoDB. The size of each message was 824 bytes. We collected and averaged the high-density lipoprotein (HDL) cholesterol levels of male patients in the age range 40–49 year, as shown in Figure 1.

```
var res = db.somecoll.mapReduce(
    map,reduce, {
        finalize: finalize,
        out:{ replace: "map_reduce_example" },
        query: {
                "HL7Message.PID.PID_8" : "M",
                "HL7Message.PID.PID_7": {"$gte": 19630401 , "$lte":19720331 },
                "HL7Message.OBX.OBX_3.OBX_3_0":  "JHDL",
            }
});

var map = function() {
    for (idx in this['HL7Message']['OBX']) {
        if (this['HL7Message']['OBX'][idx]['OBX_3']['OBX_3_0'] == "JHDL") {
            var key = "JHDL";
            var value = { sum : parseInt(this['HL7Message']['OBX'][idx]['OBX_5']), count : 1}
            emit(key,value);
            }
    }
}
```

Figure 1. Document query and map function for the message data.

Result and Discussion

Approximately 2,500 HL7 laboratory messages could be inserted into MongoDB per second. The query execution time was 501 s for the 10,000,000 messages (199,600 messages per second). Each object was 3568 bytes, and the total stored data was 36 GB. The sizes of the messages stored in MongoDB were 4.3 times larger than those of the original messages, due to conversion from the raw HL7 messages into BSON format. A performance test on a single node showed that MongoDB had sufficient performance for a university hospital system. MongoDB is a document-oriented NoSQL database, and can search schema-less documents quickly. In addition, it can be horizontally scaled by sharding, and supports distributed processing using MapReduce(3). Our results demonstrate the scalability of MongoDB, using an internal laboratory repository of healthcare data. In future work we plan to carry out a performance test on multiple nodes to confirm that our approach can meet the requirements of a nationwide repository.

References

1. Kimura M, Nakayasu K, Ohshima Y, Fujita N, Nakashima N, Jozaki H, et al. SS-MIX: A Ministry Project to Promote Standardized Healthcare Information Exchange. Methods of Information in Medicine. 2011;50(2):131.
2. M. Szeredi. File system in user space 2013. Available from: http://fuse.sourceforge.net.
3. Zuidhof R, van der Til J, editors. Comparison of MapReduce Implementations. Proceedings 8th Student Colloquium 2010; 2011.

“Structured clinical data in a schema-less database platform”

Peter Li, PhD, Mayo Clinic, Rochester, MN

Summary

There are many challenges associated with the migration of a traditional relational database to one based on more recent architecture, such as NOSQL systems. We will describe issues and solutions related to modeling, querying, and reporting for a clinical registry database.

Introduction

Traditional clinical registries are based on relational databases (RDBMS). Some are designed for specific studies with full relational capabilities, others are based on generic frameworks, often using Entity-Attribute-Value (EAV) approaches. The relative merits of each approach have been reviewed extensively. Recently, there have been large efforts on cloud-based NOSQL (key-value) systems that are conceptually based on EAV. At the same time, clinical registries are growing in size to accommodate larger consortium-based patient cohorts. This confluence presented an excellent opportunity to explore the applicability of cloud-NOSQL for clinical registries.

Methods

This project investigated the applicability of an experimental platform based on MongoDB, ElasticSearch, and Nodejs as to collect, integrate and querying the clinical data based an existing clinical registry RDBMS with over 1 million patients. This was implemented on a local computer cluster with data distributed to independently allocated nodes and segregated storage (similar to cloud platforms). This platform was used to evaluate: 1) the ease of migrating the content model from the relational system to MongoDB using available off-the-shelf and internally developed tools; 2) mapping of relational queries to ElasticSearch query strategies; and 3) query correctness vs. speed.

Results and Discussion

The clinical registry contained 557 columns spread over 23 tables. The column names have natural groupings, e.g. “first_name” and “last_name” that were used to build a flexible hierarchical model of the data per table. Of the 23 tables, 5 are “parent” tables, 15 are child tables (multi-value attributes), and 3 are extensible controlled vocabulary tables. Query speed is dependent on the sharding of the database, achieving near-linear speed-ups. While the new platform offer scalability, speed, and novel text query strategies, they do not address relationship queries, e.g. joins between independent parent tables. To perform these joins, middleware (Nodejs) solutions were implemented. Therefore, considerable effort must still be made to properly migrate a traditional system to a “schema-less” environment. In addition, the maturity of the software platform is a concern for security-minded applications. However, it is possible to develop a robust, secure, and scalable system for open-ended growth.

A Web-based Clinical Trial and Participant Registry at Einstein-Montefiore Cancer Center

Yingqin Luo[1], Kristin Fallon-Hanley[2], Bilal Piperdi[2], Parsa Mirhaji[1], Xin Zheng[1]*

[1.] Research Informatics Core (RIC), Harold & Muriel Block Institute for Clinical & Translational Research at Einstein and Montefiore, Albert Einstein College of Medicine at Yeshiva University, Bronx, NY 10461, [2.] Centralized Protocol and Data Management Unit (CPDMU), Albert Einstein Cancer Center, Montefiore Medical Center, Bronx, NY 10461

Summary We developed Einstein-Montefiore Protocol and Participant Electronic Registry (EM-PaPER), which provides a friendly user interface, employing an end-user demand driven design. EM-PaPER is a simple, secure, cost-effective, and easy-to-set-up system that enables management of the protocol registration life cycle, provides data quality control, and allows real-time reporting.

Background The Montefiore-Einstein Center for Cancer Care and the National Cancer Institute (NCI)–designated Albert Einstein Cancer Center perform research with the unifying goal of converting new findings into treatments and therapies to help cancer patients. The patient and protocol registries that the Centralized Protocol and Data Management Unit (CPDMU) has utilized since early in the Millennium bear many potential nuisances to research such as: (1) limited data quality control, (2) barriers to readily available patient and protocol information, (3) difficulties and delays with producing required reports, (4) low system availability and reliability and (5) resource intensive manual processes to abstract and collate data from multiple sources to meet business requirements. Proficient use of valuable resources is difficult due to efforts made to handle these challenges. Additionally, research investigations may suffer from the inability to retrieve complete protocol and patient information rapidly. The need for reliable and standardized electronic protocol collection and participant registration in clinical trials is apparent, especially for clinical trial office resources to more effectively manage data and studies, while consistently satisfying the reporting requirements for cancer research. In response to the growing demand for a standardized electronic application for collecting actionable information about protocols and registering Montefiore participants to oncology studies managed by the AECC CPDMU, we developed Einstein-Montefiore Protocol and Participant Electronic Registry (EM-PaPER), a clinical trial protocol and patient registration system.

Approach EM-PaPER is designed to capture both CTRP required information, and also other reporting requirements in a centralized repository by AECC CPDMU. It is developed on modern object oriented programming and relational database design technology. This application encrypts secure web interfaced application to ensure adequate security.

Results and Discussion EM-PaPER is a large-scale comprehensive and efficient system, which provides a friendly user interface, employing an end-user demand driven design. EM-PaPER is also less costly than a commercial management system. By registering and regularly updating protocols in EM-PaPER, center centers now can easily manage the life cycle of protocols (initiation through termination) through a web enabled application that provides reliable and timely information to the end users. In regards to compliance, version control of protocol documents is more effectively maintained through EM-PaPER for study coordinators and investigators. Key personnel information, including review of privileges such as consenting can be centrally accessed and managed from this system. Communication and protocol status updates is captured to allow the CPDMU to keep a central log of protocol updates. The system provides the function for the CPDMU to validate participant status and to better control data quality and patient registrations. Features built in EM-PaPER include screening information, clinicaltrials.gov registration, and participant status information amongst other data points. In addition, EM-PaPER can be easily adapted to any clinical trial office for the life cycle protocol collection and participant registration. In summary, EM-PaPER is not only a simple, secure, cost-effective, and easy-to-set-up system, but also enables management of the protocol registration life cycle, provides quality control, and allows real-time reporting.

Operationalizing Use of a Statewide Integrated Clinical Data Warehouse through a Multifaceted Educational Program

Genevieve R. Lyons[1], Katrina Fryar, MBA[1], Katherine G. Reilly[2], Jihad S. Obeid, MD[2], Christine B. Turley, MD[1]

[1]University of South Carolina, Columbia, SC. [2]Medical University of South Carolina, Charleston, SC

Abstract: *User engagement for a new multi-institutional research tool requires strategic investment of resources to address several challenges, including: identification of potential users/trainees, content, and logistics. Here we describe our multifaceted approach for training users across the state in the use of i2b2 to query our integrated Clinical Data Warehouse.*

Introduction: Health Sciences South Carolina (HSSC) is a statewide collaborative whose mission is to improve the health of all South Carolinians through applied research. Among several ongoing projects is a multi-institutional Clinical Data Warehouse (CDW). With the recent release of the CDW we have created a de-identified data mart that can be accessed using i2b2 (Informatics for Integrating Biology and the Bedside). This combination of de-identified data and query tools in i2b2 empowers users from HSSC institutions across South Carolina to access electronic clinical data in compliance with federal and institutional regulatory oversight. The HSSC CDW currently contains inpatient and outpatient records for approximately 50% of the population of the state, providing a rich bed for research.

Methods: Our goals were (1) to educate end users about the power of our CDW to provide previously unprecedented source of data for their research, and (2) to empower researchers to easily access this resource by making training accessible and understandable. To ensure successful training we used a multifaceted approach that focused on three key steps: **first**, developing a training curriculum and materials for users with a variety of backgrounds; **second**, reaching out to the member institutions to identify a local "super-user" to act as site coordinator (and eventually as a local expert); and **third**, the execution of training activities at each institution. We informed our training content by reviewing reference materials from i2b2 as well as other institutions. We developed web resources including a step-by-step text and graphical user's guide, a YouTube video demo, a FAQ page, a data dictionary, and a reference guide for ICD9 codes. In addition, we developed a presentation and that can be given in-person or via webinar. Finally, we rolled out four levels of operational support beginning with helpdesk call center for first level support. All helpdesk and support activities are tracked in Redmine.

Discussion: The successful execution of i2b2 training required a high level of communication and collaboration, especially given that participants were located across the state. By identifying the coordinator or "super-user" at each institution, we were able to work through logistical challenges and arrange several on-site sessions with videoconference access. By assembling a diverse team we succeeded in developing training content that was accurate, clear, and meaningful. The team included the HSSC Chief Medical Officer and a biostatistician, who worked with the lead systems analyst and CDW project manager to approach the challenges associated with presenting multi-institutional training to end-users who have different professional backgrounds. In particular, we carefully chose language for our materials and presentations; the training content needed to be both technically accurate and understandable for clinicians, public health researchers, statisticians, and informaticists. Also, we had to consider different scenarios in which users might be querying i2b2: for example, to determine if a large enough cohort exists for longitudinal analysis in preparation for a manuscript, grant proposal, or quality improvement. We realize that people learn best through a variety of modalities, so we made sure to utilize a variety, while also allowing learners to self-search guides and information or reach out for direct assistance.

Conclusion: As we move into the era of big data, during a time when healthcare improvement is highly emphasized, we expect that the use of data warehouses and data query tools such as i2b2 will continue to increase. The value of the tool, and of the HSSC CDW, is maximized by having a network of researchers and clinicians who are competent in using i2b2.

Acknowledgements: This work was supported by Health Sciences South Carolina (HSSC) and its member institutions and funding from The Duke Endowment and by the South Carolina Clinical & Translational Research Institute, with an academic home at the Medical University of South Carolina, through NIH Grant Numbers UL1 RR029882 and UL1 TR000062.NIH grants

The Role of Informatics Coordinator in Catalyzing Adoption of a Self-Service Integrated Data Repository Model

Tamara M. McMahon, Daniel W. Connolly, Bhargav Adagarla, Lemuel R. Waitman
Division of Medical Informatics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, Kansas

An integrated data repository was the central investment for clinical research informatics at an academic medical center. Funded upon receipt of a Clinical and Translational Science Award, the informatics coordinator serves as honest broker, educator, analyst, and has allowed informatics to manage growth in system capabilities and adoption.

Introduction: In 2010, medical informatics at the University of Kansas Medical Center (KUMC) developed the Healthcare Enterprise Repository for Ontological Narration (HERON)[1], an i2b2-based integrated data repository that provides self-service access for investigators to fully de-identified data from multiple hospital, clinic, and university data systems. KUMC faculty members and sponsored participants may access and query the data without institutional review board (IRB) approval to identify cohorts and visualize the distribution of patient populations. A data request oversight committee comprised of Kansas University Hospital, Kansas University Physicians Group, and KUMC reviews investigators' system access sponsorship and data use requests. Researchers may use and be familiar with some of the systems integrated within HERON, but very few are data experts within these systems or have a comprehensive knowledge of clinical, billing, and research systems. Reciprocally, informatics and EHR personnel understand system architecture but lack comprehensive knowledge of all clinical workflows and knowledge bases contained within these systems. As a result, self-service search poses challenges that, left unaddressed, can result in negative perceptions of informatics capabilities and failure to utilize this powerful resource. The Clinical Informatics Coordinator joined the HERON team in January 2012 to lead adoption.

Observations/Results: HERON adoption climbed steadily over the past three years though challenges continually evolve as data sources and the users expand. To address these challenges, the Clinical Informatics Coordinator's efforts in educating end users included creating training materials (instructional web pages and online video tutorials), leading a biweekly walk-in clinic for researchers, individual trainings, and coordinating a two day in-depth training workshop in August 2013. i2b2 was originally designed to facilitate translational research. The self-service model of HERON has allowed for researchers to explore other uses, such as quality improvement, an educational aid, chart review combined with HERON, refining cohorts of eligible patients who agree to be contacted for studies (Frontiers Research Participants), and augmenting their research databases by incorporating patient registry REDCap databases into HERON. HERON's target audience has expanded past researchers to include study coordinators, research assistants, hospital quality improvement, students, and administrators. The Clinical Informatics Coordinator also serves as the honest broker; acting as the liaison between the researcher and the oversight committee and providing data upon approval.

Discussion: The Clinical Informatics Coordinator position has existed 21 out of the 35 months (60%) since HERON's implementation. 15,111 searches were conducted Nov. 2010-Sep. 2013, while 75% (11,463 searches) occurred since the Clinical Informatics Coordinator joined the team in Jan. 2012. 83% (60/72) of data use requests and 70% (91/131) of sponsorship requests occurred during this time as well. Search numbers and sponsorship requests in Quarters 1-3 of 2013 already surpass totals in these areas for Quarters 1-4 of 2012 (2013 Q1-3: 6786 searches and 51 sponsorship requests; 2012 Q1-4: 4677 searches and 40 sponsorship requests) and between 20 and 40 individuals use the system monthly (with the exception of over 50 users during the month of the HERON training workshop). As noted above, HERON uses and users continue to change, and feedback from coordinator role prioritizes continuous improvement of auditing/oversight processes and streamlined methods for data extraction. The clinical informatics coordinator is seen as critical to sustaining the self-service integrated data repository model.

[1]Waitman LR, Warren JJ, Manos EL, Connolly DW. Expressing observations from electronic medical record flowsheets in an i2b2 based clinical data repository to support research and quality improvement. AMIA Annu Symp Proc. 2011;2011:1454-63.

Standards-Based Data Model for Clinical Documents and Information in the Shared Annotated Resources (ShARe) Project

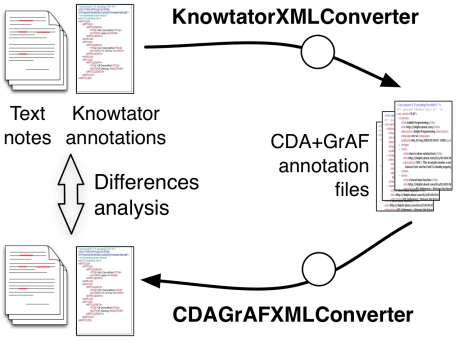
Stéphane M. Meystre, MD, PhD[1], Narong Boonsirisumpun, MS[2], Noémie Elhadad, PhD[3], Guergana Savova, PhD[4], Wendy W. Chapman, PhD[1],

[1] Department of Biomedical Informatics, [2] School of Computing, University of Utah, Salt Lake City, UT
[3] Columbia University, New York, NY [4] Children's Hospital and Harvard Medical School, Boston, MA

Abstract: *We evaluated the adequacy of a standards-based data model – CDA+GrAF – for clinical text annotations in the Shared Annotated Resources (ShARe) project, and developed tools to automatically convert annotations between the Knowtator format used in the ShARe project and CDA+GrAF. A random sample of 50 annotated notes were successfully converted back and forth, with valid and accurate annotations in both versions.*

Introduction: To support clinical research, Natural Language Processing (NLP) can be used to extract detailed information from clinical documents, but progress with applications of NLP to clinical narratives has been, and still is, significantly hindered by the lack of clinical narratives that can be easily used or shared for research applications. The Shared Annotated Resources (ShARe) project aims at alleviating this hindrance by developing de-identified and annotated sharable corpora of clinical notes.[1] To ease sharing and enable interoperability, a common information model and common terminologies are required. To answer this need, we evaluated a standards-based text annotation data model: CDA+GrAF.[2] This data model combines two existing standards (HL7 Clinical Document Architecture and ISO Graph Annotation Format) to represent all kinds of text annotations and serve as a pivot data model for annotations exchange and combination.

Methods: To evaluate the adequacy of CDA+GrAF for the representation of ShARe clinical text annotations, we focused on four objectives: 1) manually create examples of MIMIC-II clinical text annotations (according to the current ShARe use case) using the CDA+GrAF data model; 2) develop conversion tools to automatically convert text annotations from the Knowtator format used in the ShARe project to the CDA+GrAF format, and back; 3) use the conversion tools to automatically convert MIMIC-II clinical notes annotated with Knowtator in the CDA+GrAF format, and back; and 4) examine the validity of the resulting CDA+GrAF XML annotation files, and the accuracy of the annotation files translated back in Knowtator format. To work on these objectives, we randomly selected 50 MIMIC-II notes and obtained the corresponding ShARe annotations. These text annotations included multiple different categories of concepts and relations such as anatomical sites, diseases and disorders, and temporal information.



Results: The ShARe annotations are rather complex because of the various class and relation types used, but we managed to represent them faithfully in the CDA+GrAF format, without any loss of information. To easily convert Knowtator XML annotation files in the CDA+GrAF format, and back, we developed two Java conversion tools: KnowtatorXmlConverter and CDAGrAFXmlConverter. The first automatically converts Knowtator annotation files and the corresponding annotated text notes into CDA+GrAF annotation files. The second automatically converts CDA+GrAF annotation files into Knowtator annotation files and the annotated text notes. The CDA+GrAF annotation files were examined for validity and general content. The Xerces XML parser with HL7 CDA and GrAF XML schemata were used for testing, and all documents were considered "well-formed" and "valid." We also manually examined 10 of the CDA+GrAF annotation files, as visualized in a web browser using an XSL stylesheet. All ShARe annotations were correctly represented, and no errors were found.

The newly generated Knowtator annotation files and text notes were finally compared with the original Knowtator annotations files and text notes, and the analysis of differences showed that content of the former was identical (XML elements order was changed), and text files were identical. This automatic bi-directional conversion was therefore a success.

Acknowledgments: Project funded by R01GM090187 from NIGMS.

References

1. Chapman WW, Elhadad N, Savova G. Clinical NLP Annotation [Internet]. Available from: http://www.clinicalnlpannotation.org/index.php/Main_Page
2. Meystre SM, Lee S, Jung CY, Chevrier RD. Common data model for natural language processing based on two existing standard information models: CDA+GrAF. J Biomed Inform. 2012 Aug;45(4):703–10.

Using the EMR to Enhance the Clinical Research Enterprise: Protocol Validation, Feasibility, and Enrollment

Aaron Miller, PhD, Kristin A. Martinez, BA, CCRP, Steven Yale, MD
Marshfield Clinic Research Foundation, Marshfield WI 54449

Introduction

Considerable research has been done on the causes for failures or delays in obtaining clinical trial results with the most common cause due to inadequate or slow participant recruitment. To address these factors, a comprehensive feasibility process and staffing model was developed to capitalize on the use of the Marshfield Clinic electronic medical record (EMR) system.

Background

Marshfield Clinic's physicians and staff support clinical trials from a variety of sponsors (investigator initiated trials, collaborative trials developed cooperatively with external site investigators and federal and industry sponsored trials). When selecting a study for participation, historically a study underwent an ill-defined feasibility process to determine a protocol's recruitment potential, costs, and staffing needs.

Marshfield Clinic's electronic health record contains coded diagnosis, demographic, vitals, laboratory, medication and treatment data as far back as 1960. This rich datasource has been used to estimate study populations for research feasibility and directly for retrospective studies in hundreds of research projects. Challenges have occurred in using EHR data as a tool for clinical study selection when communication between clinical research staff and informatics programming break down and incorrect assumptions are made.

Methods

Effectively utilizing informatics staff to help clinical research department accurately and quickly select clinical study protocols based on retrospective EHR data requires dedicated resources to facilitate communication between staff on both teams. We have created a shared position whose primary focus is to understand the language and workflow of both the informatics group and the clinical research team. The individual selected was trained independently by both groups, and is otherwise qualified to fill a role on either team. This person is also empowered to act in a project management role, which allows them freedom to drive selected protocols when necessary.

Results

Through the use and continual refinement of this process, MC can effectively leverage resources (financial and human) to better assess clinical trial feasibility and will complete a higher percentage of trials, on time and within budgeted resources. These process improvements could extend to improved investigator initiated clinical trials and MC enhance our prospects as a preferred partner.

Real-time Federated Data Translations using Metadata-driven XQuery

Peter Mo, MS, N. Dustin Schultz, MS, Richard L. Bradshaw, MS, Ryan Butcher, MS, Ramkiran Gouripeddi, MBBS, MS, Phillip B. Warner, MS, Randy K. Madsen, BS Bernie LaSalle, BS, Julio C. Facelli, PhD
Department of Biomedical Informatics and Biomedical Informatics Core of the CCTS, University of Utah, Salt Lake City, USA

Abstract: Data federation of heterogeneous databases involves two phases of translations. The first phase is Query Translation where query criteria are translated from the harmonized data model into the disparate data models. The second phase is Result Translation where data from disparate data sources are translated back into the harmonized data model for analysis. Using the OpenFurther[1,2] data federation framework, we developed a single generic metadata-driven XQuery processor for each of these translation phases, that allows using metadata-configuration to add to the federation new data sources on-the-fly when appropriate mappings between the common model and the source exists. Currently, we are using the OMOPv2 and OpenMRS data models to demonstrate the technology, but this strategy will allow us to quickly add new data sources moving forward.

Introduction: The OpenFurther framework has been designed to address data federation challenges, and architecturally relies heavily on Representational State Transfer (REST) web services and XML. By combining XQuery's Transformation features, together with a metadata repository (MDR) and Terminology Server (TS), OpenFurther is capable of performing data translations for heterogeneous data sources using a single generic approach. This metadata-driven approach allows clinical researchers to query and view data from these otherwise disparate sources in a harmonized and meaningful way, without software code modifications for each additional unique data source.

Methods: The process of Query Translation translates end user query criteria into the structure required for each participant data source. The MDR manages translation logic, data model metadata, and harmonization mappings for each data source and is queried to determine the XQuery processing logic. The mapping properties also include data type translations logic and coded value translation logic. When coded values such as data source local terminologies need to be translated, the XQuery processor makes a request to TS and retrieves the translated coded value. Result Translation requires another critical MDR configuration. It is the XPath to the location of the attribute value in the result XML file. The result XML file structure represents each external data model and is completely different from the Query XML model. This XPath configuration allows the XQuery processor to dynamically find attribute results for translation back into the OpenFurther harmonized data model. All interactions with the MDR and TS are performed through REST web services during run-time within the OpenFurther framework.

Conclusion: Translation of federated clinical data using a metadata-driven approach provides a flexible and scalable solution for clinical research and informatics. Adding new data sources requires minimal configuration effort, if adequate mappings exist, and generally no programming changes for translations. We have successfully utilized this infrastructure for performing query and result translations against OMOPv2 and OpenMRS sample data sets. The XML translations generally take a few seconds depending on complexity of the query or results. XQuery can perform complex XML operations with minimal code. The potential drawback with XQuery is that there may be a learning curve for the unconventional FLWOR coding structure and debugging may be difficult. The XQuery translation source code, along with example input and output XML files can be viewed on the OpenFurther further-open-xquery GitHub[3] repository.

Acknowledgements: Funded by Grants D1BRH20425 from HRSA, 1UL1TR00106701 and 1TL1TR00106601 from NCRR/NCATS/NIH. Apelon, Inc. Center for High Performance Computing at University of Utah.

References

1. Bradshaw RL, Matney S, Livne OE, Bray BE, Mitchell JA, Narus SP. Architecture of a federated query engine for heterogeneous resources. *AMIA Annu Symp Proc 2009*, 70-4. PMID 20351825; PMCID: PMC2815441.
2. Livne OE, Schultz ND, Narus SP. Federated querying architecture with clinical & translational health IT application. *J Med Syst. 2011 May 3*. PMID: 21537849.
3. GitHub https://github.com/openfurther [Accessed on October 3, 2013]

Extracting Pancreatic Cancer Diagnosis and Stage from Clinical Text

Kathryn S. Nichols, MS[1], Emily Silgard, MS[2], Paul Fearn, MBA[2], Jennifer C. Yahne, MHA[3], Venu G. Pillarisetty, MD[1]
[1]University of Washington; [2]Fred Hutchinson Cancer Research Center; [3]Seattle Cancer Care Alliance; [1−3]Seattle, WA

Abstract
We have adapted natural language processing (NLP) methods to automatically extract pancreatic cancer diagnosis and staging information from electronic medical record (EMR) documents as an alternative to or facilitator of manual data abstraction. The first of its kind at the Seattle Cancer Consortium, this system shows disease progression over time and will yield a final diagnosis and stage per patient.

Introduction
Cancer research, healthcare operations, quality improvement and public health studies typically rely on resource-intensive manual data abstraction and processing to retrieve cancer diagnosis and staging information from unstructured EMR notes. Diagnosis and TNM stage (*T*umor size, *N*ode involvement and *M*etastasis) were identified as the most critical data points from the most labor-intensive sources. Our system was developed to replace or facilitate manual data abstraction of these elements from free text through document-level application of rule-based NLP methods and patient-level statistical analysis of the output. The results will serve not only as input to databases, but provide diagnosis and staging over time, valuable to other tasks such as extraction and computation of disease progression.

Materials and Methods
Patient records were extracted from the Cerner EMR through Microsoft Amalga using ICD-9 codes 157-157.9. We chose 63 oncology and procedure notes for training and 26 at random for testing. Training data annotation was completed by authors KN, ES and JY, and test data annotation by a surgical oncologist (VP). We configured pyConTextNLP to identify primary diagnosis; a set of explicit (e.g. *pT2N0M0*) or language-derived stages (e.g. *no evidence of metastasis*); whether each stage was clinical, pathological or of unspecified type; and each element's certainty (*final* or *preliminary*). Explicit clinical and pathological stages required specific modifiers (e.g. *c*, *p*, *clinical*, *pathological*) while resolution of derived stages required interpretation of text at a document level.

Results
Diagnosis was extracted at 61.5% accuracy and diagnosis certainty at 95.5%. 80% of diagnosis errors were the result of a mismatch in training and testing annotation tasks: training extracted only explicitly stated diagnoses; test annotation also used types of chemotherapy to derive histology. Under the training schema, accuracy is 92.3%. Since final/preliminary is only outputted when a stage or diagnosis element is found, final/preliminary accuracy is for the subset of elements where both system and annotator found a diagnosis or stage. Results for staging are below.

	Clinical				Pathological				Unspecified			
	T	N	M	fin./prelim.	T	N	M	fin./prelim.	T	N	M	fin./prelim.
Explicit	96.2	96.2	100	100	96.2	96.2	100	100	92.3	92.3	96.2	100
Derived	88.5	92.3	69.2	100	80.8	100	61.5	100	88.5	100	50	N/A

83.3% of M extraction errors were correct in value but not type (clinical, pathological or unspecified). Final/preliminary accuracy for unspecified derived stages is not applicable because the annotator found no unspecified elements.

Discussion
Some problems with the study were that the test and training annotation tasks were performed under slightly different guidelines, resulting in discrepancies between system configuration and evaluation. Furthermore, assessment of the task itself was not possible because test annotation was completed by a single person, ruling out inter-rater agreement. The goal of this system is to provide a final diagnosis and stage as well as the history of these data elements over time for each patient. To this end, we are smoothing anomalous elements using a statistical model of disease progression, resolving unspecified stages, and extrapolating information from events such as CT scans and surgeries. Future work will include evaluation of the annotation task, correction of persistent errors resulting from the limitations of pyConTextNLP, improving oncology and procedure note identification, developing a standard for final diagnosis and stage resolution, and outputting these elements to a database.

Leveraging a Single Instance of i2b2 Data Tables for Multiple SHRINE networks with Different Ontologies

Ayan Patel MS[1], Shawn N. Murphy MD PhD[2], Douglas S. Bell MD PhD[1]
[1]UCLA Clinical and Translational Science Institute, Los Angeles, CA [2]Partners HealthCare System, Boston, MA

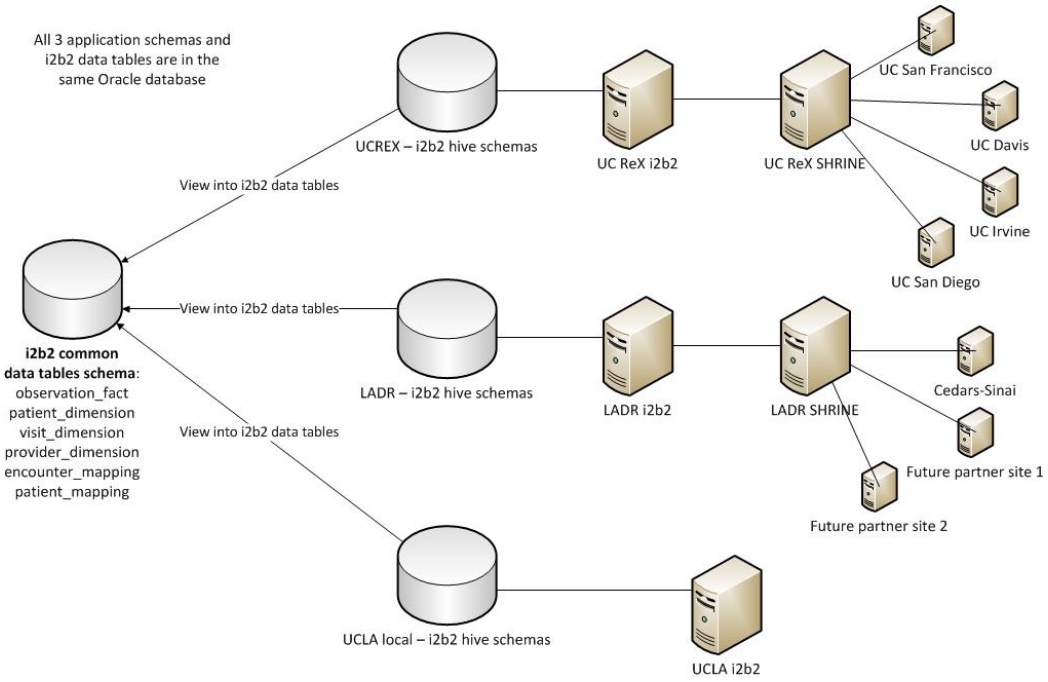
UCLA is a member of two SHRINE networks: University of California Research eXchange (UC ReX) and Los Angeles Data Resource (LADR). The UC ReX network federates i2b2 queries between the five University of California medical centers (UCLA, UCSF, UCD, UCI, UCSD), reaching 12.5 million patients. The LADR network will federate i2b2 queries among Los Angeles-area medical centers beginning with UCLA and Cedars-Sinai Medical Center reaching 6.5 million patients. Along with the two SHRINE networks, we are deploying a local i2b2 instance with 4 million patients.

The same data from UCLA will feed all three i2b2 instances. Rather than duplicating this data into three separate databases using three ETL processes, we set up one database such that the three i2b2 instances can leverage common data tables and reference the data with different ontologies.

One schema was created to contain the common i2b2 data tables (shown in the figure). The CRC cell schema of each application then referenced the common data tables with a view. All three application schemas and the new schema are located in the same Oracle database.

Due to the different governance arrangements and data harmonization priorities of the networks, each maintains its own ontology. Mapping between the common data concept codes and separate ontologies was done by adding multiple concept codes to the same concept path in the concept_dimension table, which remained within its respective application's CRC schema.

This approach reduces the complexity of maintaining multiple i2b2 instances by simplifying the ETL process and reducing disk space usage while providing the flexibility for each current and future network to upgrade SHRINE and i2b2 independently of each other and maintain separate ontologies.



Web Based Tool to Build a Parallel Corpus for English to Spanish Machine Translation

Balaji Polepalli Ramesh, MS[1], Hong Yu PhD[1,2,3]
[1]University of Massachusetts Medical School, Worcester, MA; [2]University of Massachusetts, Amherst, MA; [3]VA Massachusetts Central, Leeds, MA

Abstract

We are building a machine translation (MT) system called NoteAid$_{Spanish}$ to translate EHR from English to Spanish. As a part of the NoteAid$_{Spanish}$ development, we are building an English-Spanish parallel corpus. In this study, we report the development of a web-based annotation tool to assist human translators.

Introduction

Providing patients access their EHRs has shown to improve health care outcomes. A recent US census data reported that 17% of US population is Hispanic, of which 50% of individuals have limited English language skills. EHRs are usually written in English, which make them incomprehensible to population who do not speak English. Health professionals have been using Google Translate, which is a statistical MT system that translates text by matching patterns in millions of WWW documents[1]. It is not suitable for EHRs as they contain a large amount of domain specific jargon that does not typically appear in WWW documents. Furthermore, the Health Insurance Portability and Accountability Act of 1996 (HIPAA)[2] protects the privacy and security of individually identifiable health information so a secure MT system may be needed for US hospitals. Therefore, we are developing a MT system that automatically translates EHR notes from English to Spanish. The performance of a MT system depends on the quality of parallel corpora. Therefore, we are implementing a web-based tool to support the creation of a high quality English – Spanish parallel EHR corpus. In addition to an interface allowing a translator to upload the original text and to translate it, our tool provides the translator with additional information, including definitions, explanations and Spanish synonyms for shortened forms, complex terms and domain specific jargons that appear in EHR. This functionality may help the translator improve the translation quality.

Material and Methods

Our MT tool uses external knowledge resources. Specifically, the Unified Medical Language System Metathesaurus (UMLS) is a large multi-purpose, and multi-lingual thesaurus that contains millions of biomedical and health related concepts, their synonym names, and their relations, from over 150 vocabularies. We use UMLS to provide definitions and Spanish synonyms to domain specific jargon.

As shown in Figure 1, our MT tool displays an EHR and provides a text box for human translator to translate. If a translator highlights a term that s/he is uncertain of, a pop-up box with the definition of the term and a list of the corresponding Spanish synonyms appears. Both definition and synonyms are from the UMLS.

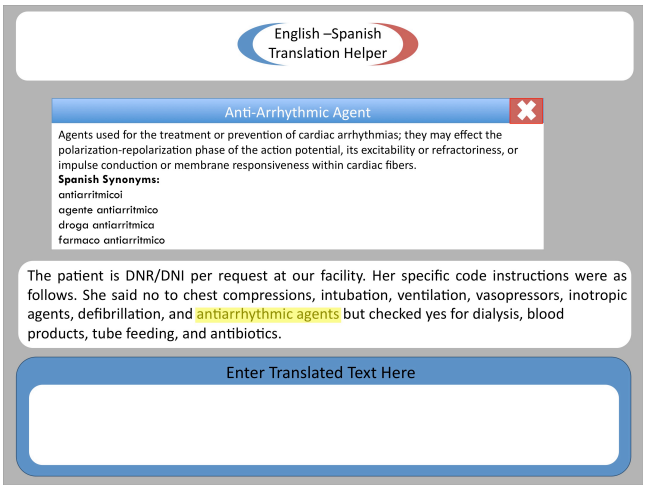


Figure 1. Screen shot of Translation Helper.

Conclusion

Our web-based MT tool assists human translators to accelerate the creation of a high quality English-Spanish parallel corpus of EHRs.

References

1. *What is Google Translate? http://translate.google.com/about/. Accessed in 2013.*
2. Health Information Privacy. <http://www.hhs.gov/ocr/privacy/index.html>
1.

Automated Identification of Comorbidities from Patient's EHR in the ER

I.V. Ramakrishnan, PhD[1], Mark Henry, MD[2], Karen Chase, BSN[2], Henry Thode, PhD[2]
[1]Computer Sc. Dept., Stony Brook University; [2]Stony Brook Medicine, Stony Brook, NY

Abstract

Comorbidities are medical conditions which are not the principal diagnosis or reason for an Emergency Room (ER) visit during a patient encounter. They affect patients' healing, survival, and length of hospitalization and is therefore important for research on predicting a patient's outcome. In the ER there is an opportunity to document patient's comorbidities. We have developed algorithms for automatic identification of comorbidities from a patient's Electronic Health Record (EHR) in the ER and automatically push them for review by the clinician. Experimental evaluation conducted in the ER at Stony Brook Medicine (SBM) demonstrates that automatic push of comorbidities substantially saves clinician's time and eliminates documentation errors and omissions.

Introduction: Comorbidities are medical conditions which are not the principal diagnosis or reason for visit during a patient encounter in the ER. Several research studies related to health services and critical care have demonstrated their role in predicting patients' outcome- e.g. see [1]. Yet, more than 80% of all patients that are seen in the ER are discharged home[2]. These patients often have many chronic health issues and comorbid conditions that are not the focus of the ER. Thus proper identification and documentation of comorbidities is important for patient care. Moreover they provide a critical data point for research on service intensity weight and risk adjusted outcome measures. But the main barrier to doing identification and documentation of comorbidities in the ER today is that automation to support this process is virtually nonexistent. In fact physicians primarily rely on costly and time-consuming manual chart review. The problem with this is that due to the voluminous amount of information in charts and the focus being on the immediate condition being treated by the physician, many of the comorbidities are either missed or not documented; in certain cases the physician may decide to ignore it as unimportant or a transient occurrence.

Method: We began a collaborative project (between Computer Scientists and ER physicians) to develop a software system that will automatically search a patient's EHR and pull comorbidities from laboratory values, vital signs monitoring, radiography reports, and other electronic records (such as medication lists) and present them for confirmation and validation by the attending physician. The system consists of a database server consisting of patients' EHRs and an application server in which the algorithms for identifying and documenting comorbidities reside. EHR data comes from several sources depending on the kinds of tests done on the patient – Labs, Radiology, Echo, etc. Comorbidities are identified by algorithmic analysis of these data. These analyses are encoded as decision making rules. These rules encode machine-processable knowledge characterizing comorbidities. Characterizations could range from being as simple as "out of range values in lab results", to "certain kinds of patterns" appearing in the textual part of reports (e.g. echo and radiology reports), to those requiring more complex reasoning. The system has been tightly integrated with the Cerner IT system used in the ER at SBM. Physicians use the Cerner interface that they are all familiar with to interact with our system for reviewing and validating the comorbidities.

Results and Conclusion: We looked at 96 charts selected at random. These were all adult patients admitted to the ER of SBM in a 1 week window. The charts were placed in three different categories based on how the comorbid conditions were identified - *physician documentation only* (no aid), *physician documentation using paper comorbidity sheet* (paper aid) and *physician documentation using electronic comorbidity diagnosis* (electronic aid). Two trained RNs in clinical documentation independently pulled and reviewed the EHRs of these selected 96 patients to manually identify and document any comorbidities associated with these patients. Their documentation served as the gold standard for measuring the accuracy of the algorithm. The miss rates were 0.3, 1.1, and 4.3 for the electronic aid group, paper aid group, and no aid group, respectively. All pairwise comparisons were statistically significant (< .001 each). The median times for reviews were 9 minutes for the RNs and 3 minutes for the physician using the electronic comorbidity diagnoses. The difference in times was statistically significant ($p < .001$).

The deployed system identifies comorbidities based exclusively on (recent and past) lab results, vital signs and body weight and height. Mining of the textual content to identify cardiothoracic-related comorbidities is underway.

References

1. Poses RM, McClish DK, Smith WR, Bakes C, Scott WE. Prediction of survival of critically ill patients by admission comorbidity. J Clin Epidemiol. 1996;49(7):743-747
2. http://www.cdc.gov/nchs/fastats/ervisits.h

Informed Consent for clinical record and Sample use in Research

Asad Rana, Adela Grando, PhD, Mona Wong, Elizabeth Bell
Division of Biomedical Informatics,
University of California, San Diego, La Jolla, CA

Introduction:

Informed consent is the process that results in the patient giving authorization for or agreeing to undergo a medical intervention. The tradition medium for informed consent forms has been paper, but paper forms have limitations. Paper forms do not make use of the vast amount of information available online today. Using electronic informed consents we can embed multimedia resources to make use of this information. Making informed consent more informative can better prepare patients to decide on participating in the medical interventions. Our team at, integrating Data for Analysis, Anonymization, and SHaring (iDASH http://idash.ucsd.edu/), made it a goal for our project to make informed consents more informative for the patient, and easier to create for the researcher.

Method:

IDASH successfully developed a tool, informed CONsent for clinical record and Sample use in Research (iCONS) 1.0, to enact an electronic informed consent at the UCSD Moores Cancer Center. The tool is being evaluated in a pilot study recruiting 160 patients. Building on the success of iCONS 1.0 we decided to develop a new tool that can be used to facilitate the creation of electronic informed consents through the use of graphical interfaces. Using Drupal as our content management service we developed iCONS 2.0. The new tool uses a open source survey software to build consent forms and upload them to Drupal (http://www.drupal.org). The user interface enables a non-programmer, i.e. medical researchers, to build their own consent forms. Furthermore, the researcher is able to embed multimedia resources within the informed consent. iCONS 2.0 was developed for tablets, and all the consent management is done online in the iCONS environment. The tool provides the option of supporting assessment tests to evaluate patient's understanding about the different sections of the informed consent. The use of assessment tests could help to know which sections of the informed consent are harder to read or difficult to understand and need further explanation. After the assessment, the patient can electronically sign the document indicating the desire to participate in the medical intervention. An encrypted PDF is generated after the completion of the informed consent form. This document can be printed and saved, but not edited. The generated PDF has the IRB stamp on every page, the date the form was signed, and the signature of the participant on the last page, as required by the IRB.

Conclusion and Future Directions:

iCONS 2.0 facilitates the development of electronic informed consents, allowing researchers to build their consent forms without requiring programming skills. It has been designed to fulfill IRB requirements, supports the assessment of participant's understanding and provides mechanisms to embed informative multimedia resources. To further evaluate the tool the Maricopa Integrated Health System in Arizona will be using iCONS 2.0 to build the electronic informed consent form for an iDASH pilot study recruiting 140 patients.

Acknowledgements: NIH for support from 1U54HL108460.

Ensuring Data Integrity, Quality, and Security in Statewide i2b2 Implementation

Katherine G. Reilly[1], Jean Craig, PhD[1], Theresia Edgar[1], Christine B. Turley, MD[2], Jihad S. Obeid, MD[1]
[1]Medical University of South Carolina, Charleston, SC, [2]University of South Carolina, Columbia, SC

Abstract: *Here we describe our implementation of a statewide clinical data warehouse containing data from three institutions. The challenge is in consolidation of data from multiple institutions, normalization of terminologies, protection of patient information, and prevention of competitive institutional identification, while providing researchers with access to meaningful and rich data.*

Introduction: Health Sciences South Carolina (HSSC) is a statewide consortium that delivers innovative solutions and enhances research capabilities across the state. HSSC's Clinical Data Warehouse (CDW) is a vital piece of the IT infrastructure and aggregates real-time clinical data from participating hospitals. Patient and encounter information is then deidentified and published to an i2b2 datamart dedicated for translational research, hypothesis testing, and cohort analysis. Compliance and security regulations are enforced throughout the CDW and i2b2 workflow and managed via HIPAA compliance, IRB protocols, secure system access, data collaboration agreements from participating institutions, data usage agreements from researchers, and audit reporting. Specifically for the HSSC i2b2 implementation, the focus is not solely on preventing patient identification, but also on preventing site identification to mitigate competitive risks between participating institutions.

Methods: Our architecture utilizes multiple projects and data schemas within i2b2: the essential "HSSC i2b2 Project" contains the full statewide set of patients and encounters, and the individual "Site i2b2 Projects" provide views of the institutional data subsets. We leverage conditional materialized views to avoid duplication of data storage among our i2b2 projects. We also limit researcher access to the HSSC Project and their home institution's Site Project.

Our i2b2 implementation began with a focus on protecting patient information. All data published to the i2b2 datamart is deidentified within our extract transform and load (ETL) process and HIPAA identifiers are removed except for dates, which are randomly shifted by up to one year. Then in order to support the cross-institutional queries for demographic and encounter concepts like Race, Hospital Service, and Financial Class, we widened our focus to leverage CDW master data management and cross-mapping for local terminologies and codesets. This cross mapping not only supplied a level of equivalence between source codes but also provided a mechanism to mask site-specific labels and prevent site identification within the HSSC i2b2 Project.

Next, querying the CDW, we identified a list of unique ICD9 codes that were only used by one of the participating institutions. Though most were coding differences, the concern was on potentially sensitive procedures that were only performed at one our member institutions. The HSSC CDW Governance Committee had each site review their list and provide feedback. To prevent site identification, sensitive codes were inactivated in the i2b2 ontology and rolled up to a more general classification in the ICD9 hierarchy.

Discussion: The resulting deidentified data set available in i2b2 is rich in detail and opens a window to a statewide population for researchers. The normalized codes help protect institutions from being identified but also apply a level of equivalence so that cross-institutional queries can be performed. For example, location specific names for cancer centers and specialty clinics were rolled up to generalized Hospital Services like Oncology and Ophthalmology for researchers to use in queries.

Conclusion: HSSC will continue to enhance i2b2 features and terminologies as new institutions are added to the CDW and as the breadth increases to include medication and laboratory data. Future expansions also include the usage of the i2b2 Modifier Dimension and SHRINE for federated queries with other warehouses.

Acknowledgements: This work was supported by Health Sciences South Carolina (HSSC) and its member institutions and funding from The Duke Endowment and by the South Carolina Clinical & Translational Research Institute, with an academic home at the Medical University of South Carolina, through NIH Grant Number UL1 TR000062.

Software for Ensuring Semantic Data Integrity: University of Florida's dchecker

keywords: semantic web, VIVO, data integrity, SPARQL, RDF

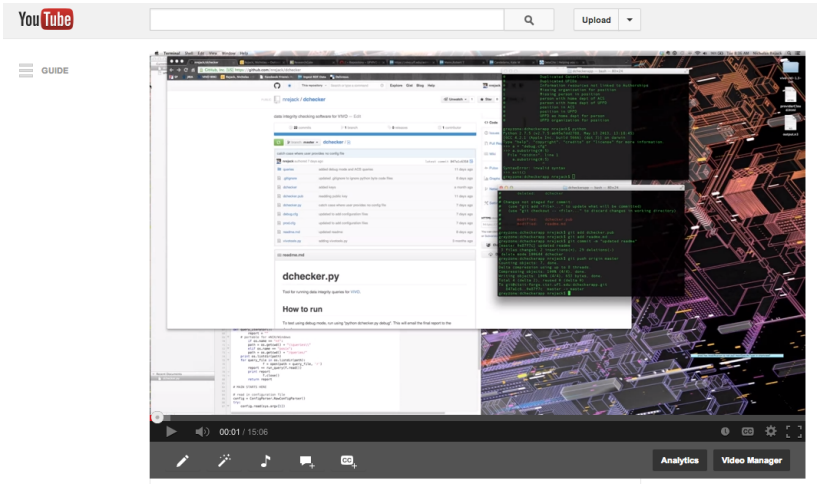
Nicholas Rejack, MS, Christopher P. Barnes, Michael Conlon, PhD, University of Florida, Gainesville, FL, USA

All data is inherently messy and requires some massaging to fit it into a manageable form, whether it is a database or a semantic system. The problem of ensuring data integrity in DBMS is well understood, but this is not the case for semantic web triple data.

Some of the constraints on data that apply to database systems have analogues in the semantic domain. For example, RBDMS restrictions that ensure referential integrity force constraints on foreign key references to valid primary keys in a parent table. Similarly, to ensure data integrity in semantic triplestores URI references within a triplestore must point to a valid, existing URI, or the links are broken. Domain integrity also has its semantic analogue: data properties defined as holding dates or string objects must not contain incorrect data types.

However by virtue of their use of agreed upon ontologies semantic systems must also maintain semantic integrity: the data must follow the agreed upon meaning defined by the ontology creators. Unique identifiers must truly be unique per individual, a property defined as a book title must not hold chapter headings, people must not also be classed as organizations, and so forth. Maintaining proper data integrity not only ensures the validity of the data as presented in human-readable format but also ensures that processes like semantic reasoning can proceed in an automated fashion.

The University of Florida has implemented VIVO, semantic software that is used for researcher networking and presenting profiles of faculty and staff members. Although the VIVO web interface enforces the domain and range restrictions of the VIVO ontology when entering data, much VIVO data entry takes place via upload of large RDF/XML files or automated harvest processes. These forms of data entry can bypass the data integrity checking and can present fragmented or incorrect data. In addition, incomplete deletion via the interface can break semantic assertions. To correct this problem, UF has developed a set of SPARQL queries and software that runs the queries on a daily basis to track and correct data integrity problems.



The software is open source and available on github (http://www.github.com/nrejack/dchecker), in addition to a tutorial video (http://www.youtube.com/watch?v=8Lz4V7HuETk). This poster will present the software and discuss ways to adapt it to other semantic web applications.

SP+ An institutional integrated data and project management system for clinical research registries

Mark E. Sakauye[1], Somchan Vuthipadadon[2], Emanuel Villa[1], Emin Kuscu[1], Chanchai McDonald[1] & Ian M. Brooks[1].
[1] Office of Biomedical Informatics, The University of Tennessee Health Science Center, Memphis, TN 38163.
[2] Institutional Research, Creighton University, Omaha, NE 68102.

Abstract:
We have developed a suite of patient, project and data management features embedded in our institutional research registry system that make running and maintaining large research registries easier and more efficient.

Introduction/Background:
The UTHSC Office of Biomedical Informatics is a core service division of the Office of Research. We consult, develop and maintain clinical trial data centers, patient registries and other tools for faculty research projects. We make use of open source (e.g. i2b2), licensed (e.g. REDCap) and internal data systems (our Slim-Prim and SP+ platforms). Increasingly we are invited to consult on multi-institutional collaborative projects. This collaborative nature leads to large and complex registries, and are thus both time-consuming and fraught with numerous technical difficulties. Indeed, by definition, a registry accrues patients/subjects. However, rarely does the number of staff working on a project increase at a proportional rate. It therefore becomes problematic for staff to both give patients the individual attention required and to keep up with accumulating data errors. Furthermore mandatory federal, local and institutional reporting can present be a major time burden. It is therefore necessary to develop tools to help staff manage these issues. We have found that although the content of each registry is unique, the concepts behind these tools can be applied to variety of registry formats. Therefore, we have developed a suite of generic staff and project management tools embedded as an integrated data system mounted on our SP+ CDMS.

Methods:
We used our Slim-Prim platform (Oracle database and CentOS web server) as the basis for developing our SP+ registry system. Our users were polled for useful features that might help them complete their tasks and perform their daily duties. These features were then independently implemented within Slim-Prim, using a combination of HTML, JavaScript, PHP and SQL. Following agile development principles, the new tools then underwent a series of rapid releases where user input was received and modifications were constantly applied until the tools perfectly satisfied their needs. Once user-satisfaction peaked, we applied concepts to all registries within the system, allowing all groups to benefit from the efforts of each project. In most cases both graphical and non-graphical (i.e. table-based) outputs are presented to users.

Results & Discussion:
It is a truth universally acknowledged, that faculty in possession of research funding, must be in want of a research database. This is indeed essential for clinical research projects seeking to accumulate electronic Protected Health Information (PHI). When such registries accumulate large numbers of patients both project management and timely data validation are a concern. The tools we have developed allow PIs and other supervisors to quickly validate data patient-by-patient or form-by-form. Viewing by patient-level, it is easy to quickly validate individual data and series' of data and to spot outliers, e.g. inverted blood pressures, aberrant A1C values in a diabetic population etc. Viewing by form-level, it becomes easy to spot abnormal clusters of data or missing values. Another problem with large registries is population management. Our other tools help alleviate concerns through interactive system-wide and individual-patient dashboards. For example, registry-wide checklists identify patients requiring follow-up. These may be combined with alerts for patients and staff if appointments or medication titration stages are missed. Our dashboard tool offers an "at-a-glance" summary of individual patient data including encounter dates, procedures and outcomes. Again, patients may even be tagged by the system based on specified flags, such that staff are alerted to provide special treatment or follow-up care. A final valuable feature provides the ability to generate a number of mandatory reports (e.g. IRB, CMS, or DSMB) in an instant. Each report is completely customizable within our system and can be exported in a variety of formats CSV, XML, SAS etc.). These tools are extensible and can be adapted to a variety different registry types, regardless of subject matter and data content. For example, we have used the same tools in cancer registries, asthma registries, and prenatal care and parent education registries. And although every report is unique, the code was written to enhance portability, with minimal modifications required to create each one. The code has the flexibility to receive a wide array of inputs and provide robust error checking, therefore ensuring consistent output.

Expressing Research Protocols in an Electronic Medical Record Using a SOAP Interface

Matthew D. Scott, Julia L. Glenn, MRA, Brian J. Kelsey, Andrew M. Cates, Royce R. Sampson, MSN, RN, CRA, Robert M. Cain, MS*, PMP, Leila M. Forney, RN, James C. Oates, MD, Jihad S. Obeid, MD

Medical University of South Carolina, Charleston, SC

***University of South Carolina, Columbia, SC**

Abstract: *We have created an interface between our home grown research portal (SPARC Request) and the Electronic Medical Record (EMR) in an effort to reduce duplicative data entry and enhance research compliance. Protocol specific data, such as study personnel and billing calendar are pushed across the interface.*

Introduction: A web-based research management system, SPARC Request, that integrates both research and routine clinical care workflows has now been in operation at the Medical University of South Carolina (MUSC) since March 2012. Concurrently, MUSC was transitioning its outpatient EMR to Epic. As the primary portal for all research services on campus, SPARC Request needs to interface with MUSC's EMR, Epic, to ensure research billing compliance, data consistency across systems, and patient safety. As part of the planning for full Epic rollout due in the near future for patient registration and inpatient areas, we have worked closely with the Epic team and subject matter experts to integrate research services into the Epic workflow.

Methods: A research Epic team was assembled. Using agile software development practices and pair-programming, and through close collaboration with Epic, a unidirectional interface has been built allowing protocol, principal investigator, associated study personnel, and billing calendar information to be pushed from SPARC Request to Epic. Epic supports data exchange web services and a SOAP (Simple Object Access Protocol) interface using an Interconnect server. As a result SOAP was used to push data in real-time into Epic. By examining the data structures of both SPARC Request and Epic we were able to determine the maximum amount of relevant information that was common to both, and then the SPARC Request team began creating the interface from scratch. Despite having limited access to a testing server, and work-in-progress API documentation, a module was built in Ruby to generate an XML SOAP message, and that module seamlessly integrated into the main Ruby on Rails application.

Results and Discussion: The interface has been completed and is undergoing extensive evaluation in a testing environment. Initial feedback from our subject matter experts has been positive. The interface is scheduled to go live in the near future with the completion of the Epic rollout in July 2014.

The Interface is able to push all of the information we need to translate the billing protocol to Epic. Beginning with basic protocol information to create a study in Epic, we are able to associate the study with the relevant users by passing along their SPARC Request identifiers. Moreover, the entire billing protocol for that study is transferred, including relevant dates, subject and visit counts, individual services requested, and billing modifiers on individual instances of a service being provided.

The system does have its limitations. It would be preferable for processes to be automated. For example, currently upon submission of a service request in SPARC Request our Epic admins have to be notified to insure that all relevant users already exist in Epic before the actual push to Epic can take place. With functionality enhancements to Epic's research module expected in the next major release, perhaps it will be possible to cut out that step by populating those users programmatically.

Those future enhancements to Epic's research module will essentially dictate the future of the interface. As the system is increasingly able to accept additional information about protocols and service requests, and building on the same methods and lessons learned from building this initial interface, the SPARC Request-Epic interface will be continuously enhanced to push as much information as possible into Epic to reduce duplication of effort on the part of researchers.

Conclusion: Coordination between research systems and EMR's is critical for a successful research enterprise. We believe this interface will have significant return on investment by improving billing compliance, patient safety and reduced redundancy in data entry.

Acknowledgements: This work was supported by the South Carolina Clinical & Translational Research Institute, with an academic home at the Medical University of South Carolina, through NIH Grant Number UL1 TR000062.

Characterization of Clinical Data Elements for Secondary Use in a Comprehensive Cancer Center

Emily T. Silgard, MS[1], Paul A. Fearn, MBA[1,2],

Kathryn S. Nichols, MS[1], Jennifer P. Tran[1], Angelica Omaiye[1], Nishant Velagapudi[1]

[1]Fred Hutchinson Cancer Research Center (FHCRC); [2]University of Washington (UW), Seattle, WA

Abstract

As foundational research to support enterprise data integration and automation, we characterized clinical data elements across the FHCRC/UW Cancer Consortium, where development of numerous research and operational databases have led to inefficient, redundant manual data abstraction and complex, varied information architectures. This research provides a strong case for employing methods to ease the burden of manual data abstraction from unstructured text, enabling clinical and research staff to retrieve and use clinical information more efficiently, thereby improving healthcare operations and advancing cancer research.

Introduction

At the FHCRC/UW Cancer Consortium there are approximately 5,000 new cancer patients a year and 150,000 historical patients. It is time- and resource-intensive for researchers and clinicians to acquire clinical data in the right forms for each use. Our objective was to normalize names for conceptually identical data elements across disparate databases, determine opportunities for data architecture simplification and automation by tracing elements to their original source systems, assess whether elements could be patient-reported or derived from other elements, and identify candidates for information extraction with natural language processing (NLP). The results of this analysis are informing the design of an NLP and computational pipeline to replace or support manual processes and the creation of a generalized clinical data model for cancer to allow for more efficient information retrieval.

Materials and Methods

Materials consisted of 18 data dictionaries from existing and prospective systems in 28 subgroups within 13 major disease groups, comprising 14,220 metadata elements across 1114 tables. In a largely manual analysis spanning 10 weeks and 600 person-hours, we filtered out elements not directly related to clinical histories, normalized field names, and identified source systems and structure of data elements. This iterative process involved over 66 discussions with respective clinicians, researchers, data managers and developers who confirmed that alleviating the burden of manual data abstraction was critical. Data elements were analyzed first by disease and then merged to characterize them across all diseases by each feature (i.e. source, structure, patient reported, computed).

Results

From the original elements, 7565 were unique concepts. Without administrative elements (e.g. primary and foreign keys, timestamps), 6685 elements remained, of which 4000 were unique. 15% of the 4000 elements could be computed, 15% could be patient-reported and 65% were manually abstracted from unstructured text and therefore candidates for NLP. Of elements from unstructured text, 50% came from clinical notes, 20% from pathology reports, 15% from surgery notes, and 5% from radiology reports. The resulting normalized data elements predictably followed a Zipfian distribution, where a handful of elements (such as demographics, diagnosis and staging) occurred in numerous databases while the vast majority (e.g. specific lab tests and procedure details), occurred only once. These accounted for 90% of the final clinical elements. Of the singletons, 20% were derived from other elements, 10% could be patient reported, and 33% pertained to specific procedures or therapies.

Discussion

Much of the complexity could not be assessed at the metadata level. Source of elements may vary from patient to patient, so results were estimated assuming that the initial diagnosis and all treatment occurred within the Consortium. Based on tumor registry statistics, less than a third of Consortium patients are diagnosed and treated here, so 35% is a best case estimate for elements from structured sources. We also used a conservative definition of computed; elements were considered computed if deriving them required another element. Future work will include using this data to architect a data model and NLP pipeline for a Consortium-wide integrated data repository, with links to their associated biospecimens, studies and molecular data. Although clinically relevant data elements within cancer care are fairly standard and the need to access data within free text is ubiquitous, data systems vary widely, so it's difficult to say how well these results would carry over to other institutions.

CIELHO: A Platform for Distributable Research Analytics

William E. Stephens; David Ervin; Philip R. O. Payne, Ph.D. The Ohio State University, Department of Biomedical Informatics, Columbus, OH

Summary

Distributed research networks allow researchers to perform distributed queries for the purposes of patient-centered research [1]. Expanding upon this use case, we discuss CIELHO, a software platform intended to enable the authoring, sharing and execution of analytical bundles that enable complex, distributed analysis and development of analytical workflows.

Introduction

The Collaborative Informatics Environment for Learning on Health Outcomes (CIELHO) project is currently being implemented to support the need for sharing analytic methods and tools that support research. It is being developed in collaboration with the Academy Health Electronic Data Methods (EDM) Forum under funding from the Agency for Healthcare Research and Quality (AHRQ). The primary purpose of this platform is to provide members of the research community with access to an open-source/standards "app store" for data analysis and software sharing through which they can access others' applications, contribute back their own, and build upon each other's contributions.

Background

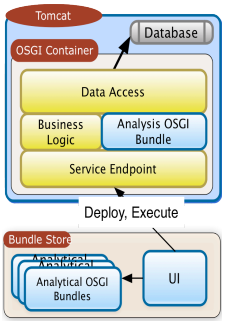
The current distributed query capabilities of patient-centered research networks have limitations. CIELHO enables researchers to extend or assemble modules in the areas of Comparative Effectiveness Research (CER), Patient-Centered Outcomes Research (PCOR), and Quality Improvement (QI). Additionally, this work is complementary to efforts to create open access to study data and algorithms by providing greater transparency and reproducibility around the data and analytics used to generate results.

Methods

Central to our engineering effort is the implementation of a set of software tools that will enable authoring, sharing and execution of analytical bundles against heterogeneous and multi-dimensional data sets. Based upon an initial and iterative design process involving key stakeholders, we have identified several key components as being necessary to support the aforementioned "app store" functionality. Tooling is provided to encapsulate Java-based analytical software into Open Services Gateway Initiative (OSGI) bundles. These bundles are signed with TRIAD x.509 user certificates in order to create a verifiable fabric of trust among the shared software in the research community. Researchers upload these signed bundles and contribute associated metadata, including specification of the required clinical data structure, to our analytical "bundle store" web application in order to make them available to the research community. Once uploaded, these bundles are available for discovery and deployment into available "wrapper" applications that allow CIELHO-compliant bundles to be invoked against a data structure that is made accessible via a CIELHO-compatible controller application(s). Our initial "wrapper" web service, which will be an extension of our TRIAD services produces in association with SAFTINet[1], provides support to securely deploy a bundle and perform analysis against an Observational Medical Outcomes Partnership (OMOP) data structure.

Results

Our development team has established an architecture that enables the implementation, packaging, sharing and execution of trusted analytical algorithms against distributed data sources that extends beyond the simple querying available through current frameworks.



Discussion

CIELHO serves as the initial implementation of a framework that enables the patient-centered research community to perform detailed analytical computation against a selection of distributed data stores, thus increasing the transparency and reproducibility of research results while maintaining secure control over their shared institutional data.

References

1. Schilling, Lisa M.; et al (2013) "Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) Technology Infrastructure for a Distributed Data Network," *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 1: Iss. 1, Article 11., DOI: 10.13063/2327-9214.1027, Available at: http://repository.academyhealth.org/egems/vol1/iss1/11

Acknowledgement

The authors would like to acknowledge the EDM Forum for funding this project as well as the contributions of Erin Holve, Jonathan Nebeker and Michael Kahn during the conceptualization of the CIELHO platform.

Intermountain iDiscover™: Mobile, compliant, secure clinical trials research software

David P. Taylor, PhD, Angela Schwab, MS, CGC, Matthew Ebert, MS, Jason Gagner, MBA, Kira Wagner, BS, Edinardo Potrich, MS, Peter J. Haug, MD

Intermountain Healthcare, Salt Lake City, Utah

Summary

iDiscover is a clinical trials research management system internally developed to meet investigator-initiated research needs of a large healthcare delivery system. It provides innovative form authoring and delivery tools, auditing, integration with existing systems, and mobile access. The software has been a model for the advantages of use-case driven development.

Introduction and Background

Investigator-initiated research studies are common at Intermountain Healthcare but paper-based workflows for participant enrollment, consent, investigator approvals, and research data collection (e.g., case report forms) have been the standard of practice. A number of commercial and free-of-charge software packages are available for clinical trials management and data collection. The decision concerning system acquisition was based on factors such as the ability to support existing workflows, cost, implementation time, customizability, and ability to integrate with existing data sources and systems. After surveying the marketplace a decision was made to build a research management system locally with a focus on investigator-initiated clinical trials.

Methods

The software was produced through a collaboration between the Homer Warner Center for Informatics Research and the Intermountain Heart Institute although other clinical research groups are also now represented. A product steering committee and a user group were created to facilitate gathering frequent input from research coordinators, principal investigators, and other research personnel.

Results

iDiscover includes a native iOS app and a web browser client. The iOS app (for use with an iPad today) is used for situations where mobility is key (e.g., participant enrollment, consent, case report forms). The browser client also includes study/form authoring tools and other study administration features. No data are stored on an iOS device or client machine, but transferred securely using SSL encryption and stored centrally. Role-based access is provided for particular studies and individual forms. iDiscover has the ability to access data from existing Intermountain systems (e.g., patient lookup and EHR access for clinical results) to reduce the need for redundant data entry and potential errors.

Discussion

iDiscover supports the specific needs of Intermountain researchers in ways that would be difficult for externally available tools to meet. It provides a highly flexible forms authoring tool with a library of predefined templates, branching logic, calculated fields, and sub-forms accompanied by a full audit trail history. Study participants can review IRB consent forms and sign electronically on an iPad. Study documents can be routed for approval electronically (for instance to the principal investigator). Data are exportable from the system in various formats and the ability exists to de-identify any fields if desired.

However, a key value of this system is as a platform for future enhancements to our research environment. We are currently working on the following enhancements: (1) Automatically identifying potential research participants through rule-based and probabilistic methods, (2) Facilitating the participant consent process through the use of multimedia, (3) More use of existing participant data (e.g., demographic, clinical) from the EHR, (4) Supporting research document management, (5) Facilitating the scheduling of research visits and workflows within and between visits, (6) Integrating with research financial systems, (7) Producing mobile apps for researchers summarizing current studies, inclusion/exclusion criteria, and study protocols, (8) Expanding dashboards and reporting, (9) Supporting multi-center trials (i.e., externalizing the software), and (10) Creating tools for participants to access study information and provide data outside the office setting. Additionally we plan to support standard medical terminologies, integration of the data within the Intermountain Clinical Data Repository and Enterprise Data Warehouse, and incorporation of workflow management tools to assist with sophisticated, multistep research protocols.

Governance of a Multi-institutional Integrated Clinical Data Warehouse

Christine B. Turley, M.D[1], Katrina Fryar-Riley[1], Jihad S. Obeid M.D.[2]
University of South Carolina, Columbia SC[1], Medical University of South Carolina, Charleston, SC[2]

Abstract: *Governance of a multi-institutional integrated Clinical Data Warehouse is complex and must incorporate privacy and confidentiality issues at the patient level, as well as competitive practices and business intelligence at the institutional level. We describe a multi-institutional governance structure spanning South Carolina that incorporates both regulatory and corporate interests.*

Introduction: Health Sciences South Carolina (HSSC) is a statewide collaborative designed to facilitate clinical research success in SC with a goal of improving the health of South Carolinians. It is comprised of the three research intensive universities (University of South Carolina, Medical University of South Carolina and Clemson University) and the 4 largest health systems in the state (Palmetto Health, Medical University Hospital Authority, Greenville Health System and Spartanburg Regional Health System). A cornerstone component of the collaborative has been the development of an integrated Clinical Data Warehouse (CDW) which is contained in a Limited Liability Corporation, Health Sciences Health Improvement (HSHI). The CDW combines Health Level 7 (HL7) data messages as well as clinical data from component local data repositories or data warehouses. A multi-institutional integrated CDW is governed by federal regulations for privacy and confidentiality, as well as being subject to research oversight through the Office of Human Research Protection (OHRP), the Common Rule and HIPAA privacy and security rules. Additionally, each institution has business interests that create competitive positions around clinical programs.

Methods: The initial development phase of the CDW was governed by a Memorandum of Understanding between the HSSC member institutions and HSSC. As we prepared to allow access to de-identified data marts, a more extensive Data Collaboration Agreement (DCA) was developed and put into place and IRB approval was obtained across all organizations. This DCA has provided an overarching framework for the development of an active governance program. Features of this governance program include: 1) compliance with federal and state law; 2) compliance with the DCA signed by all parties; 3) provision for the security of all data; 4) creation of de-identified data marts which contain multi-institutional data in a manner that neither unmasks a patient or an institution; and 5) creates a process for researchers to get early, appropriate assistance to assure that data requests are both appropriate and consider the guidance supplied by the DCA.

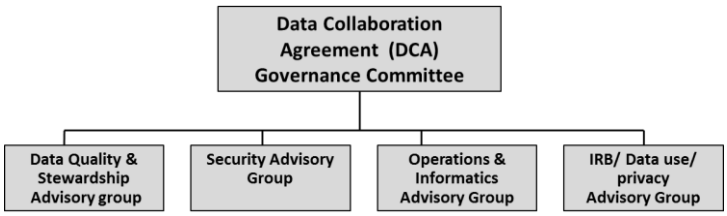


Figure 1. HSSC governance structure.

The Governance Committee (GC) is comprised of designated members from each HSSC member institution. This GC is informed by advisory groups convened to address and provide expert advice on topics including: a) Data Quality and Stewardship, b) Security, c) Operations and Informatics, and d) IRB, Data Use Privacy (Figure 1).

Discussion: The Governance of an integrated CDW that crosses multiple institutions is complicated by matters that extend beyond regulatory matters. These institutions are competing for both research funds and clinical care and clear structures to manage these issues is foundational. Important topics undertaken for review by each advisory group with the appropriate expertise are developed into recommendations and presented to the GC for evaluation. Factors integral to successful execution of this model have included building on a history of collaboration on clinical improvement projects, infrastructure development projects[2], and a commitment to the overarching collaborative.

Conclusion: We share a roadmap for developing governance of an integrated CDW and use models that protect interests across disparate enterprises. Comparison to components of governance models in place and/or recommended at other institutions are considered. Governance of data across these entities presents interesting challenges and opportunities in the changing health care landscape.

References

1. Murphy, SN, Weber, G, Mendis, M, Gainer, V, Chueh, HC, Churchill, S, Kohane, I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17:124-30.
2. Sanderson IC, Obeid JS, Madathil KC, Gerken K, Fryar K, Rugg D, Alstad CE, Alexander R, Brady KT,Gramopadhye AK, Moskowitz J. Managing clinical research permissions electronically: A novel approach to enhancing recruitment and managing consents. *Clin Trials*. 2013 Jun 19. PMID: 23785065.

Explorations with Distributional Semantic Models for Ad hoc Medical Concept Search

Jay Urbain, PhD[1,2], Paul Knudson, MD[2,3], Brady Taylor, BS[2,3], Glen Bushee, MS[2,3]

[1]Milwaukee School of Eng., [2]CTSI of SE Wis., [3]Medical College of Wis., Milwaukee, WI

Summary

There are many applications in translational research that require the identification of fine-grained ad hoc concepts and their relations from disparate repositories of text. Such concepts may be of a specialized type for a specific information need, or defined through an ad hoc knowledge discovery process. We explore a solution to this problem with an online ranked retrieval and extraction framework based on a distributional semantic space model. In the proposed model, contextual evidence of concepts and relation instances is integrated across words, phrases, sentences, and documents. By using a novel high dimensional indexing strategy, semantic, syntactic, and lexical evidence can be efficiently aggregated, allowing the system to collectively learn and discover fine-grained entities at query time.

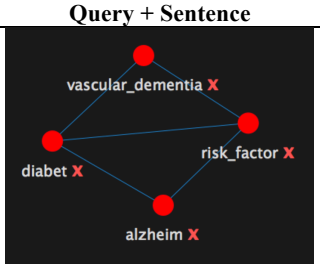
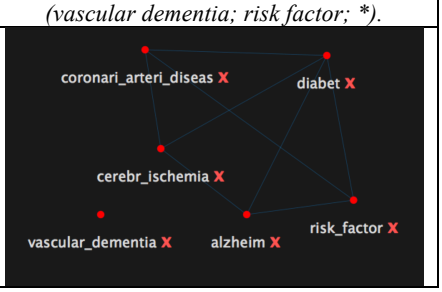
Background

In an ad hoc open domain search setting, specific concept and relation types may not be known, so there is a need for measuring the semantically similarity of candidate concepts and relations. These needs are in contrast to the extractions provided by traditional information extraction (IE) systems that are constrained to a predetermined set of targeted concepts and relations where extractions are defined in a knowledge base, embodied in a set of rules, or defined via extraction models trained from specialized training data. Targeted extractions from such systems may be too general, unavailable, or domain specific to adapt to ad hoc information needs.

Methods

- A novel dimensional index to represent a Vector Space Model (VSM) of distributional statistics. The index facilitates efficient OLAP style SQL queries for aggregating query-time statistics of semantically, lexically, and syntactically related concepts, relations, and terms.
- New unsupervised (and semi-supervised) retrieval and extraction models integrating state-of-the-art models of semantic, lexical, and syntactic evidence to identify semantically related medical concepts and relations.
- Fully functional, scalable prototype based on Java/MySQL/JSP and Amazon Web Services.

Table 1. (a) Concept-relation search result for query: *Diabetes related to Alzheimer's*. (b) Graph of query and sentence result. (c) Concept-relation graph search results for query: (*vascular dementia; risk factor;* *).

a) Retrieved Sentence with concepts & relational dependencies	**b) Concept-relation graph: Query + Sentence**	**c) Semantic similarity graph: query:** *(vascular dementia; risk factor; *).*
Diabetes is a risk factor for vascular dementia. *Dependency relations: (concept1; relation 1, 2,...;concept2)* diabetes; ; risk_factor risk_factor; *for*; vascular_dementia diabetes; *risk_factor_for*; vascular_dementia		

Results and Discussion

Each component model was evaluated on benchmark data. Unsupervised aggregate *entity relations* where evaluated against i2b2 2010 Challenge data set and open-domain web data and received results average precision of 0.88, which is similar to systems using predetermined, supervised learning models.

Electro**n**ic **T**ailored **I**nfographics for **C**ommunity **E**ngagement, **E**ducation, and **E**mpowerment (EnTICE[3])

Mark Velez[1], Michael E. Bales, PhD[1], Adriana Arcia, PhD, RN[2], Suzanne Bakken, PhD, RN[1,2]
[1]Department of Biomedical Informatics and [2]School of Nursing, Columbia University, New York, NY

Introduction and Background: Through the Washington Heights/Inwood Informatics Infrastructure for Comparative Effectiveness Research (WICER) project, we collected data about social determinants of health and health behaviors on more than 6,000 community residents. Consistent with the overall goal of WICER which is to understand the health of the primarily Latino community in order to improve it, we have undertaken a set of activities to return data to the community in a manner that is acceptable, comprehensible, and motivates self-management of health concerns. We developed a set of infographics and conducted focus groups with 97 community members to assess preferences and acceptability of our designs, which are intended to be comprehensible across individuals with varying levels of health literacy. In this abstract, we describe the functional requirements and framework for **E**lectro**n**ic **T**ailored **I**nfographics for **C**ommunity **E**ngagement, **E**ducation, and **E**mpowerment (EnTICE[3]), an adaptable, reusable, and generalizable approach for creating tailored infographics based upon survey data.

Methods: Achievement of related objectives in WICER and its successor, WICER 4 U, is integral to EnTICE[3]. These include understanding needs of stakeholders (e.g., survey participants, community-based organizations, researchers) and extending the infrastructure that generates data for EnTICE[3] and the community-oriented website (GetHealthyHeights.org) on which the infographics will be available for viewing. To generate requirements and inform system design, we adopted an iterative approach, which includes use case analysis, prototyping and evaluation.

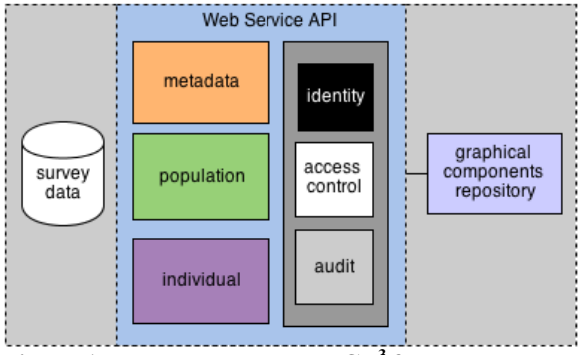


Figure 1. The proposed EnTICE[3] framework

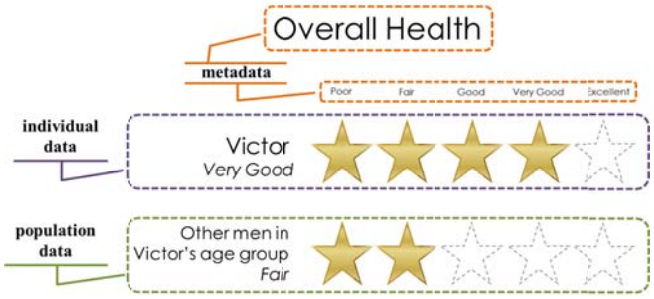


Figure 2. Sample infographic annotated with data sources

Results and Discussion: We identified a broad range of stakeholders with distinct and evolving information needs. Requirements elucidated include the provision of multiple views of the dataset, interoperability with existing and forthcoming systems, and reusable, adaptive user interface components. The proposed framework, depicted in Figure 1, consists of a service-oriented architecture that securely mediates survey data access and a repository of web-based graphical components. A sample infographic, annotated with the data sources of the graphical elements is shown in Figure 2. Preliminary evaluation demonstrated the system's ability to incorporate participant-specific attributes in rendering infographics. We plan to use EnTICE[3] to study the efficacy of infographics in communicating health status and risks to participants. Subsequently, we will use the framework to produce tailored infographics for engagement of survey participants on GetHealthyHeights.org. Given that tailored information is more likely to motivate behavior change, we consider this approach to be an essential component of our strategies to improve the health of Washington Heights/Inwood.

Acknowledgments: This research is supported by R01HS019853, R01HS022961, NYS Department of Economic Development NYSTAR (C090157). Dr. Arcia is supported by T32NR007969.

A Cross-sectional Design for Evaluation of Clinical Decision Support

Kavishwar B. Wagholikar, MBBS, PhD[1], Ronald Hankey, MS, MBA[2], Robert A. Greenes, MD, PhD[4], Hongfang Liu, PhD[1], Rajeev Chaudhry, MBBS MPH[3]
[1]Biomedical Statistics and Informatics, [2]Population Management Systems Group, [3]Primary Care Internal Medicine, Mayo Clinic Rochester, MN; [4]Biomedical Informatics, Arizona State University, and Mayo Clinic, Scottsdale, Arizona

Abstract. *As an alternative to the cross-validation design for evaluation of clinical decision support (CDS) systems, we propose a method for three-way comparison of decisions made by i) providers assisted by a CDS system, ii) un-assisted providers and iii) the CDS system itself. We present the preliminary results from a case study.*

Background. The relatively short timeframes for the 'meaningful-use' stages are a significant challenge for performing effective validations of clinical decision support (CDS) systems. The traditional design of cross-validating a CDS system by comparing the system's decisions with a set of gold standard decisions is time as well as labor intensive. We present an alternative methodology that is more efficient and feasible.

Methods. The proposed method involves a snapshot study entailing a three-way comparison of decisions made by i) CDS assisted providers, ii) un-assisted providers and iii) the system itself. A plot of the ratios of the concordance between the three groups (expressed as log ratios in figure 1), provides a characterization of the system's accuracy in terms of precision and sensitivity, and also projects the potential benefits of the system for improving the clinical practice. In figure 1, the X and Y axis are log (A/S) and log (U/A) respectively where,

A: Total number of cases where providers assisted by the CDS system recommended the test.
U: Total number of cases where unassisted providers recommended the test
S: Total number of cases where the CDS system recommended the test

We applied this approach in a case study to evaluate a CDS system for preventive care that is deployed in primary care practice at Mayo Clinic Rochester. We invited 10 providers to perform chart review for 30 patients to decide on the recommendation for 43 different types of preventive tests/services recommended by the system. The patients were randomly distributed among the physicians, such that the chart review for each patient was performed by two different physicians – one with CDS assistance and the other without assistance. 18 of the 43 tests were recommended for at least 10% of cases by either physician groups, and were retained for analysis. We computed the above parameters for each of these preventive services to visualize the system performance (figure 1) in comparison with the CDS assisted and unassisted groups of the providers.

Results and Discussion. Results indicate that lipid panel, low-density lipoprotein (LDL) and glucose tolerance test are under-recommended by the CDS system, but are over-prescribed by the care providers. The CDS system was under-sensitive for TDAP immunization. Overall, the methodology appears to provide useful insights for improving the CDS systems, but further research for validation and optimization of the proposed methodology is necessary.

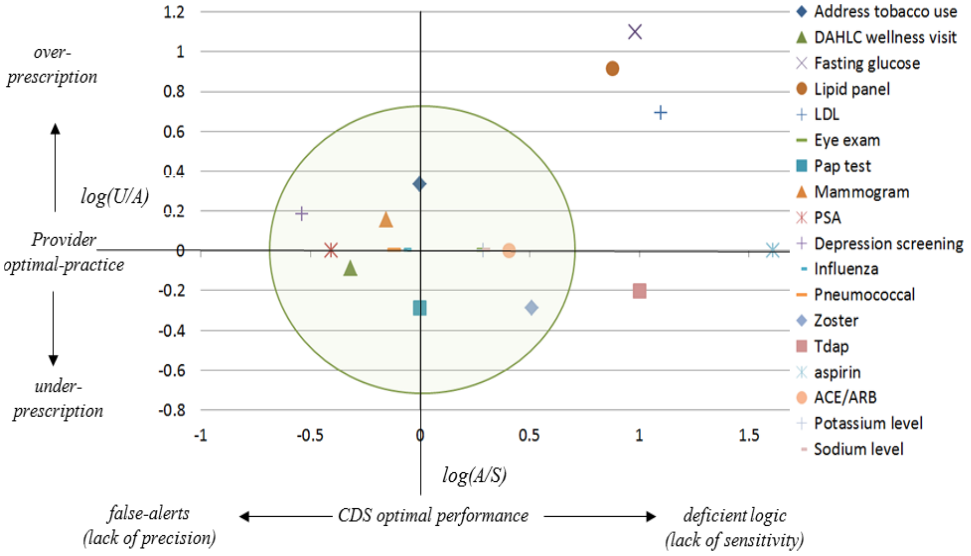


Figure 1. *The Y–axis categorizes the system utility and the X-axis categorizes the CDS system performance. Area outside the central region indicates that the system has either deficient accuracy or has high utility for improving clinical practice. 18 of the 43 tests were recommended for at least 10% of cases by either physician groups.*

A Typology of Research Paradigms and Design Principles: Clinical Trials, Disease Registries, and Electronic Health Records

Rebecca Wilgus, RN, MSN, Shelley A. Rusincovitch, Charlotte L. Nelson, MS, Monica M. Horvath, PhD, April E. Chester, MSA, MSW, Benjamin Neely, MS, Paramita Saha-Chaudhuri, PhD, Rachel L. Richesson, PhD, James E. Tcheng, MD, Robert M. Califf, MD

Duke Medicine, Durham, North Carolina

Abstract: *The need to advance clinical science using highly optimized, efficient research practices is widely recognized. The proliferation of electronic health records and heightened awareness of the potential value gained by leveraging this data are catalysts driving the transformation of existing clinical research methodologies. Understanding the nuances in purpose, workflow, and data is necessary to most effectively design and apply new research methods. This typology of three research paradigms illuminates important factors for stakeholders to acknowledge and incorporate into the design of research initiatives that span the continuum.*

Background: Research methods such as randomized clinical trials (RCT), disease registries, and analysis of data generated through healthcare service delivery are sources of evidence that advance science and evolve medical practice. The expectation to elevate the standards of care, the proliferation of electronic health data, and pressing need to reduce research costs and time are fostering the rapid emergence of innovative, applied research methods such as pragmatic clinical trials, population surveillance strategies, and predictive analytics. This rapid emergence presents opportunities and challenges for investigators to optimize their research designs to more efficiently generate evidence and synthesize knowledge by extending data use beyond its normative context.

Methods: Beginning with the consensus that, while a wide spectrum of clinical research designs exists, three central research paradigms emerge: 1) Clinical Trials; 2) Disease Registries; and 3) Health Care Delivery. Purpose, workflows and data were characterized. A multi-disciplinary team performed a comparative analysis and documented a typology of these paradigms and model of the workflows. Inherent similarities and differences were explored and described.

Results: Commonalities and differences in purpose, workflow and data have broad implications for the design of research initiatives. Highlights of the analysis are outlined in Table 1. Distinctions between data collection practices, data quality management, and visit scheduling emerge as pivotal, especially when modeled as temporally-dependent design activities and workflows.

	Clinical Trial Data	Disease Registry Data	Health Care Data
Primary Use Case	Phase III trial of an investigational agent	Performance improvement and outcomes assessment	Observational research without direct patient intervention
Data Optimization for Research Question	HIGH: Intervention and data collection are designed to address a specific, tightly-controlled research question	MODERATE: Research questions are modified to the available care delivery processes	LOW: Research questions must be modified to fit available observations and data
Data Collection	• Content is targeted to answer specific research questions • Clinical judgment and interpretation required to report the requested data • Missing data and quality issues are carefully managed • Data elements are tightly defined • Visit schedule and measurement methods and techniques are tightly prescribed • Visits occur within finite period of time • Consistently reported by trained research personnel	• Content is targeted toward assessment of clinical outcomes or performance measures • Clinical judgment and interpretation are required to abstract responses from data reported through an EHR • Missing data and quality issues are generally addressed, but may not be corrected at source (EHR) • Data elements are tightly defined • Visits occur regularly, infrequently but over longer to infinite periods of time • Consistently reported by trained specialists	• Transactional and encounter-specific • Content is driven by standards of care and operational and billing requirements • Reflects very general to highly specialized care delivery processes • Missing data and data quality issues are common • Data element definitions vary within and across EHR systems • Measurement methods and techniques vary widely within standard of care • Visit schedule and frequency are highly individualized • Reported by multi-disciplinary health care providers
Generalizability	Dependent on sample size; generally limited to a population similar to study participants	Represents a cross section of patients with specific disease condition	Heterogeneous, clinically diverse population reflective of actual clinical practices and the "real world"
Follow-up Information	Limited to finite study period	Longitudinal while patient is followed by a participating institution	Longitudinal while patient receives care within a given health system

Table 1. Selected dimensions from the typology of research paradigms.

Discussion: Appropriate and effective research design shortens time from hypothesis generation to dissemination of knowledge, makes the most appropriate use of available resources and supports the evolution of evidence-based practice. By recognizing both the opportunities for optimization and inherent limitations of each research paradigm, investigators can most appropriately leverage clinical and operational data to design efficient and cost-effective research projects that ultimately improve healthcare delivery processes and elevate the standard of care. The model produced and implications for research design identified through this work will be conveyed in the presentation.

A New Corpus for Clinical Events with Change of State

Meliha Yetisgen, PhD[1,2], Prescott Klassen, MS[2], Lucy Vanderwende, PhD[3,1], Fei Xia, PhD[2,1]
[1]Biomedical and Health Informatics, [2]Department of Linguistics, University of Washington, Seattle, WA; [3]Microsoft Research, Redmond, WA

Abstract

Understanding the event structure of sentences and the whole documents is an important step in being able to extract meaningful information from text. Our task is the identification of critical illness phenotypes, specifically pneumonia, from clinical narratives. To capture those phenotypes, it is important to identify the change of state for events, in particular events that measure and compare multiple states across time. In this abstract, we describe a corpus annotated for events with change of state information. Our corpus is comprised of chest x-ray reports, where we find many descriptions of change of state comparing the volume and density of the lungs and surrounding areas.

Introduction

The narrative accompanying chest X-rays contains a wealth of information that is used to assess the health of a patient. X-rays are obviously a single snapshot in time, but the report narrative often makes either explicit or, more often, implicit reference to a previous X-ray. In this way, the sequence of X-ray reports is used not only to assess a patient's health at a moment in time but also to monitor change. Critical illness phenotypes such as pneumonia are consensus-defined diseases, which means that the diagnosis is typically established by human inspection of the data rather than by means of a test. We are in the process of developing a phenotype detection system for pneumonia. In order to train and evaluate the system, we asked medical experts to annotate the X-ray report with phenotype labels and to highlight the text snippets in the report that supported the phenotype labeling. Analysis of the text snippets revealed that most of these snippets mention a change of state or lack of a change of state (i.e., persistent state). We created a corpus from the 1008 highlighted text snippets and annotated them for events with change of state.

Annotation

In our annotation schema[1], an event in our corpus is represented as a (loc, attr, val, cos, ref) tuple, where *loc* is the anatomical location (e.g., "lung"), *attr* is an attribute of the location that the event is about (e.g., "density"), val is a possible value for the attribute (e.g., "clear"), *cos* indicates the change of state for the attribute value compared to some previous report (e.g., "unchanged"), and *ref* is a link to the report(s) that change of state is compared to (e.g., "prior examination"). Not all the fields in the tuple are required to be present in an event. When a field is absent, either it can be inferred from the context or it is unspecified. Figure 1 includes example chest x-ray report sentences with event annotations. More detailed information can be found in our annotation guideline downloadable from UW-BioNLP website (http://depts.washington.edu/bionlp/index.html).

Three annotators who are graduate students annotated events in the 1008 snippets. 100 of the snippets were annotated by all three annotators. The inter-rater agreement at the tuple level was 0.85 macro f-score and 0.89 micro f-score.

(1) The *lungs* are clear.
(lungs, <density>, clear, -, -)
(2) *Lungs*: No focal opacities.
(lung … focal, opacities, no, -, -)
(3) The *chest* is **otherwise unchanged**.
(chest, -, -, otherwise unchanged, -)
(4) *Left base* opacity has **increased** and *right based* opacity **persists** which could represent atelectasis, aspiration, or pneumonia.
(left base, opacity, -, increased, -)
(right base,opacity, -, persists, -)
(5) **Since the prior examination** *lung volumes* had **diminished**.
(lung, volumes, -, diminished, prior examination)

Figure 1. Example snippets with event annotations.

Future Work

Our ultimate goal is to train a statistical event detection approach and use it in phenotype detection and other NLP systems to monitor patients' medical conditions over time and prompt physicians with early warning, expecting that this will improve patient health care quality while reducing the overall cost of health care.

Acknowledgements

Microsoft Research Connections, University of Washington Research Royalty Fund, UL1TR000423

References

1. Vanderwende L, Xia F, Yetisgen-Yildiz M. Annotating Change of State for Clinical Events. Proceedings of the 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation Workshop of NAACL'2013, 2013.

On the Bayesian Derivation of a Treatment-based Cancer Ontology

Michael Gao[1], Jeremy Warner MD, MS[2,3], Peter Yang MD[4], Gil Alterovitz PhD[1,5,6]

1 Center for Biomedical Informatics, Harvard Medical School, Boston, MA; 2 Department of Medicine, Division of Hematology & Oncology, Vanderbilt University, Nashville, TN; 3 Department of Biomedical Informatics, Vanderbilt University, Nashville, TN; 4 Department of Medicine, Division of Hematology/Oncology, Massachusetts General Hospital, Harvard Medical School, Boston, MA; 5 Children's Hospital Informatics Program, Harvard Medical School, Boston, MA; 6 Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA

1 Abstract

Traditional cancer classifications are primarily based on anatomical locations. As knowledge is heavily compartmentalized in the oncological specialties, discovering new targets for existing drugs (drug inference) can take years. Furthermore, our lack of understanding of the mechanisms underlying drug efficacy sometimes undercuts the effectiveness of genetic approaches to drug inference. This study tackles the twin problems of cancer reclassification and drug inference by constructing a global cancer ontology inductively from treatment regimens. A topological abstraction algorithm was performed on the bipartite graph of drugs and cancers to highlight important edges, and a Bayesian algorithm was then applied to determine a new treatment-based classification of cancer, producing 6 highly significant clusters ($p < 0.05$), confirmed by Fisher's exact test and enrichment analyses. Edge probabilities derived from its drug inference routine matched real edge frequencies ($R^2 \approx 0.96$). Drug inference results were reinforced by the identification of relevant published Phase II and III clinical trials, and the drug inference routine differentiated between high- and low-likelihood targets ($p < 0.05$). This novel treatment-based ontology has the potential to reorganize cancer research and provide powerful tools for drug inference using global patterns of drug efficacy.

2 Introduction

Two major issues in oncology are rational cancer reclassification and the efficient inference of the effectiveness of drugs against cancers other than their initial target, which we will refer to henceforth as the drug inference problem.

Throughout the history of oncology, the discipline has been split into subfields based primarily on the anatomic location of cancer. The current partitioning of the field of oncology has led to the compartmentalization of knowledge. Even within the same subfield, there is a tendency to split between the study of untreated patients and of relapsed patients, driven primarily by the exacting needs of clinical drug testing and approval.

Currently, many drugs are studied for one specific cancer and one specific context only immediately after their development, decreasing their impact considerably. As a result, discovering additional treatment contexts for a new drug can take a long time. For example, the drug imatinib was first found effective in chronic myelogenous leukemia (CML) and gastrointestinal stromal tumors (GISTs) in 2002 [1, 2]. Despite the fact that the drug was known to target c-KIT and that it has been known that certain melanomas harbor c-KIT mutations since 2005 [3], imatinib was not shown to be effective for c-KIT mutated melanoma until 2011 [4]. This long process demonstrates the need for a global solution for drug inference.

The development of large-scale biological databases has enabled researchers to explore patterns shared by cancer subtypes and target certain protein pathways crucial to the development of cancer for treatment via inhibitors. The ontological methods developed in recent years in computational genomics provide new tools for such an analysis. In a bioinformatics context, ontology is defined as the study of hierarchical classifications generated from biological data that can be used to test biological hypotheses. Recent developments in bioinformatics sustained by genomic sequencing and ontological methods have attempted to provide computational solutions to the above two problems. These solutions have adopted an approach involving the construction of models for cancers based on specific biological mechanisms such as oncogenes, protein pathways, or gene functionality [5, 6, 7].

Such an approach based on biological mechanisms is powerful in directing future cancer research, but further investigations following its guidance sometimes cannot find supportive empirical outcomes. For example, after a highly significant single nucleotide polymorphism (SNP) was found in the v-Raf murine sarcoma viral oncogene homolog B1 (BRAF) gene in melanoma patients, the drug vemurafenib was developed to target the relevant protein and led to great improvements in the treatment of melanoma [8]. The BRAF SNP was later found to be present in a significant proportion of colorectal cancers, but the use of vemurafenib in colorectal contexts has largely failed [9, 10]. Since the current literature still cannot explain many common phenomena that have a high impact on treatment efficacy, including tumor-host interactions [11], drug efflux mechanisms [12], and other indirect mediators of drug resistance, approaches to drug inference that focus on a limited range of biological mechanisms are vulnerable to such challenges.

This study provides a unified solution to the problems of insufficient cross-specialty communication and of drug inference in cancer research by developing a novel cancer-context and drug ontology. Differently from previous approaches that attempt to pinpoint the biological causes of cancer, this approach is defined by a systematic, large-scale, quantitative analysis of the existing database of cancer treatment regimens. In contrast to previous cancer studies which build a biological model of cancer first and then infer drug efficacy accordingly, this study takes an inductive approach using data mining techniques to form a standardized cancer treatment database, then constructing the aggregate pattern of cancer subtypes. Furthermore, previous approaches consider a few key biological mechanisms that lead to cancer, whereas we black-box the currently unknown, complicated biological processes underlying cancer by using the effectiveness of existing treatment regimens as an indicator of their joint impact. An additional contribution of our approach is the global

nature of the meta-analysis. Instead of comparing the mutations of only a few cancer subtypes at a time to infer drug efficacy on new targets, we compute the likelihood that any existing drug can be applied to any new target in the sorted clusters of cancers, allowing a more global drug inference study. Thus, our drug inference algorithm is capable of magnifying the effectiveness of existing cancer treatments.

To cluster cancers in a clinically meaningful way, the clustering algorithm needs to meet several criteria. We need a hypergraph to represent the relationships between cancer contexts and drugs, with edges corresponding to the set of cancer contexts treated by each drug. An effective algorithm needs to use edge weights, as certain treatments have more evidence of efficacy than others. The algorithm should determine the optimal number of clusters if the clustering is not hierarchical. Furthermore, the clusters should contribute to a probabilistic framework for drug inference, and each cluster should have an associated treatment profile. A soft clustering algorithm is optimal, as we should be able to capture uncertainty about whether a certain cancer should be assigned to one cluster or another. And finally, the algorithm should devalue unlikely treatment profiles in order to find the most plausible clustering.

Bayesian hypergraph clustering methods have addressed each of these concerns. The algorithm can incorporate edge weights by linearly weighting each edge-wise calculation, ensuring that high-evidence treatments have more impact on the clustering. A Bayesian method naturally eliminates nodes from extraneous clusters, inferring the optimal number of clusters as a byproduct of maximizing the information score (negative log likelihood) of the clustering [13]. Bayesian methods also provide probabilities for cluster assignments and edge generation between clusters and nodes, enabling a probabilistic solution to the drug inference problem using a soft clustering approach. Additionally, some Bayesian algorithms calculate rational priors for the parameters, discounting unlikely treatment profiles [14].

By supporting such a cluster analysis, a Bayesian algorithm can provide a novel treatment-based ontology of cancers that addresses both the cancer reclassification problem and the drug inference problem simultaneously. We can then test the accuracy of the drug inference results against the existing literature and use this accuracy as a metric to confirm the quality of the reclassification.

3 Methods

The dataset used in this study is from the cancer regimen online knowledge management system of HemOnc.org (`http://www.hemonc.org`), developed by Warner and Yang. It contains over 160 drugs, 480 regimens, and 50 cancers that are linked in a network. Some medications included in this dataset were excluded from this analysis as they are supportive. These included growth factors, bone modifying agents, and other supportive medications. Steroids were retained if and only if they were an integral part of a chemotherapy regimen.

Cancers themselves are mainly classified by anatomic location and treatment context in this dataset. We reclassified treatment regimens, when possible, between the previously treated and untreated contexts. When regimens were a mixture of first-line and second-line treatments, they were split into treated and untreated sub-contexts.

Most regimens in this dataset had already been classified into primary and secondary contexts, but a minority required further classification. One method used to distinguish primary treatments from second-line treatments was a naive classification algorithm run on the abstracts of PubMed papers associated with the treatments. For example, if "adjuvant" was found in the title of a paper, the associated treatment was considered a treatment for the untreated (primary) context. Slightly under 50% of all PubMed papers referenced by the database that were not already classified in the metadata were successfully classified using this algorithm.

Edges are key in determining the optimal clustering of a network. The important edges in the network are first separated out by using the Alterovitz principal component analysis-based (PCA) algorithm [15]. This provides the function of preserving the most important treatments for each cancer, and thus increasing the specificity of the treatment database. The Vazquez Bayesian clustering algorithm [13] was then applied to the hypergraph generated by the adjacency sets of vertices in the abstracted graph. Details of the computational process are provided in the Appendix. Clinical interpretations were assigned to clusters based on commonalities in treatment strategy, and then Fisher's exact test was applied to determine the treatment information enrichment provided by the Bayesian clustering algorithm.

4 Results

The extraction of the most important features of G_{DDs} resulted in a reduction from 936 edges to 589 edges. As each cancer has many minor treatments that interfere with optimal clustering, the extraction of important features by the topological abstraction algorithm directly enabled the Bayesian clustering.

After abstraction, from the Bayesian algorithm, 19 clusters were found. The clusters ranged in size from 1 to 21; the 14 clusters with at least two cancers are shown in Table 1, in which r/r diseases represent second-line treatment contexts, or cases in which patients had previously been treated.

With the confidences calculated for these cluster assignments and the θ-values calculated by the clustering algorithm for hyperedge incidence probabilities, extrapolated confidence in treatment efficacy was calculated using Equation (1). These clusters could generally be qualitatively characterized by a set of shared treatments without any consideration of the hyperparameters, confirming their clinical value. These treatments are briefly characterized in Table 1.

Clinical interpretations were then assigned to these clusters by finding commonalities in the treatment strategies of the cancers. The principal treatment patterns found in these interpretations were then tested for statistical significance, using the Fisher's exact test calculation of treatment information enrichment. Six of the 14 clusters that had more than one cancer were found to have high statistical

Table 1: Computed cancer classification and its clinical relevance.

No.	Shared treatment	Members of cancer cluster	p-value
1	Nucleoside analogs	CML r/r, AML r/r, APL, CNS NHL untreated, ALL untreated, T-NHL untreated	$5.95 \cdot 10^{-6}$
2	Platinums	Ovarian r/r, HL r/r, SCLC r/r, Sarcoma untreated, NSCLC untreated	$8.21 \cdot 10^{-2}$
3	Platinums / taxanes	Breast r/r, Bladder untreated, Cervical untreated, H&N untreated, Esophagus untreated, Breast HER2+ untreated	$1.63 \cdot 10^{-3}$
4	Immunotherapy	Melanoma untreated, Renal r/r	$3.93 \cdot 10^{-5}$
5	5FU / Folinic acid	Pancreatic untreated, Esophagus r/r, Gastric untreated, Colon r/r, Cervical r/r, Rectal untreated, HCC r/r	$1.22 \cdot 10^{-4}$
6	R-CHOP	HIV NHL untreated, MCL untreated, Aggressive NHL, FL r/r, Thymoma untreated	$7.81 \cdot 10^{-9}$
7	MTOR inhibitors	Renal untreated, ALL r/r, MCL r/r	$1.60 \cdot 10^{-2}$
8	–	CML untreated, Brain, NET r/r	–
9	–	AML untreated, CLL	–
10	–	FL untreated, HL untreated	–
11	–	CNS NHL r/r, T-NHL r/r	–
12	–	MDS untreated, Melanoma r/r	–
13	–	Anal untreated, Bone r/r, NET untreated, MPD untreated, HCC untreated	–
14	–	Thymoma r/r, Amyloid, MZL r/r	–

significance ($p < 0.05$), with the lowest having $p < 10^{-8}$, as shown in Table 1. A clinical interpretation was found for one other cluster. Treatment strategies for other clusters were not evaluated due to the lack of data in the treatment database about those clusters. This significance demonstrates that the algorithm not only constructed clinically meaningful clusters, but was able to optimally partition the set of cancers among clusters to maximize clinical information across all clusters.

Our clustering technique also provides a powerful solution to the drug inference problem. Ranking cancer-treatment pairs according to descending order of computed likelihoods, we found that input edges, edges that had already been placed in the database, represented the bulk of the high-likelihood edges, as shown in Figure 1, demonstrating that our Bayesian model achieved a reasonably close fit to the existing data.

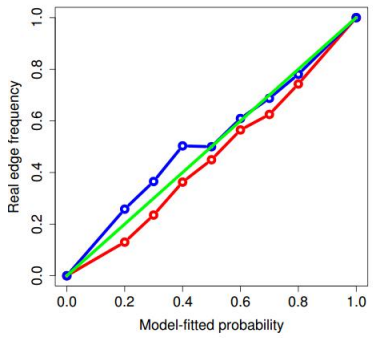


Figure 1: Model-fitted probability vs. real edge frequency

The lower bound curve, at point $0.1k$, denotes the real occurrence frequency of edges with computed confidence between $0.1(k-1)$ and $0.1k$, whereas the upper bound curve represents that of edges with confidence between $0.1k - 0.05$ and $0.1k + 0.05$. Taking the points (0.1, 0.1), (0.2, 0.2), ..., (1.0, 1.0) as predicted values, we obtained $R^2 = 0.958$ for the lower bound line and $R^2 = 0.960$ for the upper bound line. The figure thus demonstrates that the Bayesian computed probabilities correspond quite clearly to the database's edge incidence probabilities, showing that our clustering-derived incidence model closely approximates the original hypergraph, passing a basic sanity test.

For each possible edge, the probability of the edge occurring according to the Bayesian model was calculated. To measure the probabilistic approach's performance on the drug inference problem, we took its ten highest-confidence newly inferred edges and reviewed the literature. These edges, their model-derived probabilities, and associated clinical trial references are shown in Table 2.

Table 2: Inferred treatment recommendations and confidence levels

Drug	Cancer subtype	Probability	Reference
Fluorouracil	HCC r/r	0.88	[16]
Dexamethasone	Bladder untreated	0.83	[17]
Carboplatin	Sarcoma untreated	0.79	[18]
Cisplatin	Sarcoma untreated	0.79	[19]
Gemcitabine	Sarcoma untreated	0.79	[20]
Folinic acid	HCC r/r	0.74	[21]
Temozolomide	CML untreated	0.72	
Methotrexate	APL r/r	0.70	[22]
Cytarabine	T-NHL untreated	0.70	[23]
Cytarabine	CML r/r	0.70	[24]

As a control group, the bottom ten inferred edges, picked from above 1% confidence, were also considered in the literature review. As shown in the last column of Table 2, we found that while eight of the ten high-confidence edges had mentions in Phase II/III literature and the other two were being studied in non-clinical contexts, only two of the ten low-confidence edges had mentions in Phase II/III literature. Indeed, one of the low-confidence pairs was found to have a negative instead of positive relationship between treatment and cancer. A significant difference ($p = 0.023$) therefore exists between the high-confidence and low-confidence edges inferred by the clustering algorithm, and this demonstrates that our algorithm is capable of differentiating between promising treatments and treatments that are likely to fail. In short, although these treatments had already been discovered by other investigators, we were able to infer them from a database that did not contain them.

Thus, our algorithm was able to discern the hidden structure of G_{DDs} and has contributed both to the solution of the drug inference problem and the cancer reclassification problem, with the latter being confirmed both quantitatively by the drug inference results and qualitatively by inspection of common treatments by cluster and by inspection of the relationships between cancers in the same cluster.

5 Discussion and Conclusion

The results demonstrate the strength of the treatment-based Bayesian clustering algorithm and provide a simultaneous solution to the cancer reclassification problem and the drug inference problem. While previous works represent the biologically-motivated approach to oncology, this study represents a new direction of global network meta-analysis on existing treatment regimen data.

Aside from the traditional anatomical classification of cancer, recent reclassification studies have been performed on a local scale. For example, breast cancer was split into many different subtypes [25].

Our meta-analysis takes a holistic, inductive approach, using clinical efficacy data as its main variable, which reflects the entirety of all biological mechanisms behind cancer. Our treatment-based ontology sheds light on several rational cancer reclassifications. For example, in the traditional anatomy-based model, renal cancer and melanoma are deemed to be unrelated. According to our treatment-based ontology, however, they belong to the same cluster because they are similarly treated diseases, commonly found to coexist [26], and commonly treated by the same specialists. Such a reclassification will reorganize oncological knowledge, as the traditional divisions among anatomical locations will be replaced by the patterns of shared treatment efficacies.

Although our approach does not open the black box of the biology behind cancer, it does consider these mechanisms indirectly, through treatment efficacy. Instead of considering etiology directly, we first reclassify cancers based on treatment efficacy, and then suggest further investigations into similarities in the underlying biology. For example, in the case of renal and melanoma cancers (cluster 5), our new reclassification suggests that specialists working on the two cancers may jointly investigate new approaches to immunotherapy. Similarly, thymoma, which is treated primarily by CHOP regimens, unexpectedly occurred in the B-cell non-Hodgkin's lymphoma clusters, which shares those regimens but is not related anatomically. Thus, similarities in the underlying etiologies of cancers in this cluster may be jointly investigated by their respective specialists. All of these examples had high statistical significance for treatment information enrichment, as shown by Table 1. The significance of these biological validations is three-fold. These similarities represent a preliminary biomedical validation of our reclassification of cancers, and the statistical significance of the information enrichment lends credibility to our approach to drug inference. Furthermore, our findings suggest that the Bayesian algorithm, applied to a more complete database and augmented with further biological information, may be capable of quickly identifying new commonalities between the etiologies of cancers, which would merit further investigation.

Previous efforts at drug inference have focused on SNPs and oncogenes, among other genetics-motivated biological mechanisms. As the development of drugs such as vemurafenib has shown, the challenge to this approach is that it considers only a small number of the biological mechanisms that jointly determine drug efficacy and is therefore often ineffective as a solution to the drug inference problem. As a result, the potential impact of new chemotherapy drugs is limited to one cancer until oncologists slowly begin to experimentally apply them to other cancers.

Though drug inference has often been performed on a local scale, comparing the genetic and molecular profiles of two cancers at a

time, we take a global approach to the problem of drug inference, unifying it with a cancer reclassification model. Our focus on treatment efficacy rather than its main causal factors enabled this study to map global similarities between cancer subtypes by first finding clusters, then computing probabilities for the efficacy of repurposed drugs in a unified model. The effectiveness of the Bayesian algorithm at the drug inference problem is demonstrated by its differentiation between likely and unlikely treatment recommendations. As the last column of Table 2 demonstrates, likely treatment candidates had a high correlation with appearances of Phase II/III clinical trials in the literature. Unlikely treatment candidates, in contrast, had a much lower frequency of mentions in the literature. Thus, the drug inference extrapolations were clinically relevant. This suggests that a global model for cancer reclassification may be able to simultaneously address the drug inference problem.

The inductive statistical method of treatment efficacy analysis adopted by this study suggests a new way of discovering cancer knowledge by analyzing the rapidly accruing digital data on cancer treatments. At one level, it offers an alternative to models built on key genetic profiles and biological mechanisms. At another level, however, it also complements these biological models by providing a new way to organize cancer specialties and infer drug efficacy according to the treatment-based clusters identified by our statistical algorithm.

However, this paper represents only the preliminary step in exploiting our Bayesian network analysis approach in drug inference. Though we have confirmed the efficacy of the Bayesian approach in analyzing incomplete cancer-drug databases, the next step is applying the approach to a more complete database, which would allow us to make novel treatment recommendations.

In our work towards solving the cancer reclassification and drug inference problems, we have identified several important questions that must be addressed to perfect the power of our clustering algorithms and to extend the impact of our findings.

One of the potential improvements to our approach is the inclusion of more data in our meta-analysis. One possible approach is to adjust the Bayesian model; new hyperpriors can be designed for the distribution of treatment effectiveness. Another way to make more information available to the clustering algorithm would be to provide more information on absolute efficacy, in the form of negative edges. If a certain drug was found in a study to be completely ineffective against a particular cancer, a separate ineffective hyperedge should be created to include this information in clustering considerations. Although hemonc.org did not have negative information as it was meant to be a treatment guideline database, a more comprehensive database would yield better clusters. In addition to efficacy data, the inclusion of direct biological mechanisms in our inductive approach could provide a powerful syncretic method that may solve the problem of discovering new subtypes of cancer as well.

Furthermore, the possibility remains that the unit of our clustering, the cancer treatment context, is not the best disease unit to use. For example, gastrointestinal stromal tumors are included in the sarcoma contexts, but are treated differently from most sarcomas. Dividing cancers into contexts in some other way may provide more information or better clusters.

In addition to addressing these problems in clinical oncology, cancer treatment network analysis can also yield better ways to organize other processes relating to cancer care, such as drug production. Cancer drug shortages are a major problem in cancer treatment [27]. We can track sudden bursts in publications in particular clusters using a network analysis algorithm running on a cancer treatment network. Thus, it may be possible to predict drug shortages in the future, ensuring that production can be increased before the demand spike.

Analytical models of cancers based on biological etiologies have characterized cancer meta-analysis and drug inference thus far. The inductive, global, treatment-based approach outlined in this paper directly analyzes treatment efficacy data to provide a unified solution to both cancer reclassification and drug inference. Our combination of topological abstraction and Bayesian techniques was effective at elucidating the structure of the cancer treatment network provided by the regimen database, discovering hidden commonalities between cancers whose validity is confirmed by our Fisher's exact test p-values ($p < 0.05$), and suggesting new directions of research using treatment patterns found in the database. Its drug inference extrapolation routine, which black-boxes biological mechanisms by considering final treatment efficacy instead of partial sets of biological data, also yielded positive results ($p < 0.05$) in drug inference, as shown by our review of clinical literature. This Bayesian approach is also flexible enough to accept new forms of treatment efficacy data to further improve its impact.

References

[1] H. Kantarjian, C. Sawyers, A. Hochhaus, F. Guilhot, C. Schiffer, C. Gambacorti-Passerini, D. Niederwieser, D. Resta, R. Capdeville, U. Zoellner, et al. Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *New England Journal of Medicine*, 346(9):645–652, 2002.

[2] G. D. Demetri, M. von Mehren, C. D. Blanke, A. D. Van den Abbeele, B. Eisenberg, P. J. Roberts, M. C. Heinrich, D. A. Tuveson, S. Singer, M. Janicek, et al. Efficacy and safety of imatinib mesylate in advanced gastrointestinal stromal tumors. *New England Journal of Medicine*, 347(7):472–480, 2002.

[3] C. Willmore-Payne, J. A. Holden, S. Tripp, and L. J. Layfield. Human malignant melanoma: detection of braf-and c-kit–activating mutations by high-resolution amplicon melting analysis. *Human pathology*, 36(5):486–493, 2005.

[4] J. Guo, L. Si, Y. Kong, K. T. Flaherty, X. Xu, Y. Zhu, C. L. Corless, L. Li, H. Li, X. Sheng, et al. Phase ii, open-label, single-arm trial of imatinib mesylate in patients with metastatic melanoma harboring c-kit mutation or amplification. *Journal of Clinical Oncology*, 29(21):2904–2909, 2011.

[5] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[6] S. Kim, M. Kon, C. DeLisi, et al. Pathway-based classification of cancer subtypes. *Biology direct*, 7(1):1–22, 2012.

[7] K.-B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje. Multiple svm-rfe for gene selection in cancer classification with expression data. *NanoBioscience, IEEE Transactions on*, 4(3):228–234, 2005.

[8] P. B. Chapman, A. Hauschild, and C. Robert. Improved survival with vemurafenib in melanoma with braf v600e mutation. *N Engl J Med*, 364:2507–2516, 2011.

[9] K. Affolter, W. Samowitz, S. Tripp, and M. P. Bronner. Braf v600e mutation detection by immunohistochemistry in colorectal carcinoma. *Genes, chromosomes & cancer*, 52:748–752, 2013.

[10] E. C. Stites. The response of cancers to braf inhibition underscores the importance of cancer systems biology. *Science signaling*, 5:46, 2013.

[11] S. Ogino, J. Galon, C. S. Fuchs, and G. Dranoff. Cancer immunologyanalysis of host and tumor factors for personalized medicine. *Nature Reviews Clinical Oncology*, 8(12):711–719, 2011.

[12] R. K. Vadlapatla, A. D. Vadlapudi, D. Pal, and A. K. Mitra. Mechanisms of Drug Resistance in Cancer Chemotherapy: Coordinated Role and Regulation of Efflux Transporters and Metabolizing Enzymes. *Curr. Pharm. Des.*, Jul 2013.

[13] A. Vazquez. Finding hypergraph communities: a bayesian approach and variational solution. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(07):P07006, 2009.

[14] E. T. Jaynes. Prior probabilities. *Systems Science and Cybernetics, IEEE Transactions on*, 4(3):227–241, 1968.

[15] G. Alterovitz and M. F. Ramoni. Discovering biological guilds through topological abstraction. In *AMIA Annual Symposium proceedings*, volume 2006, page 1. American Medical Informatics Association, 2006.

[16] C. Porta, M. Moroni, G. Nastasi, and G. Arcangeli. 5-fluorouracil and d, l-leucovorin calcium are active to treat unresectable hepatocellular carcinoma patients: preliminary results of a phase ii study. *Oncology*, 52(6):487–491, 1995.

[17] A. L. Gruver-Yates and J. A. Cidlowski. Tissue-specific actions of glucocorticoids on apoptosis: A double-edged sword. *Cells*, 2(2):202–223, 2013.

[18] D. Goldstein, B. Cheuvart, D. Trump, M. Shiraki, R. Comis, D. Tormey, J. Harris, and E. Borden. Phase ii trial of carboplatin in soft-tissue sarcoma. *American journal of clinical oncology*, 13(5):420–423, 1990.

[19] A. Waddell, A. Davis, H. Ahn, J. Wunder, M. Blackstein, and R. Bell. Doxorubicin-cisplatin chemotherapy for high-grade nonosteogenic sarcoma of bone. comparison of treatment and control groups. *Canadian journal of surgery. Journal canadien de chirurgie*, 42(3):190, 1999.

[20] S. Okuno, J. Edmonson, M. Mahoney, J. C. Buckner, S. Frytak, and E. Galanis. Phase ii trial of gemcitabine in advanced sarcomas. *Cancer*, 94(12):3225–3229, 2002.

[21] G. Di Lorenzo, A. Rea, C. Carlomagno, S. Pepe, G. Palmieri, R. Labianca, A. Chirianni, A. De Stefano, V. Esposito, S. De Placido, et al. Activity and safety of pegylated liposomal doxorubicin, 5-fluorouracil and folinic acid in inoperable hepatocellular carcinoma: a phase ii study. 2007.

[22] S. Nagai, T. Takahashi, and M. Kurokawa. Risk-adapted maintenance therapy for acute promyelocytic leukemia. *Journal of Clinical Oncology*, 28(2):e21–e21, 2010.

[23] S. J. Kim, K. Kim, Y. Park, B. S. Kim, J. Huh, Y. H. Ko, K. Park, C. Suh, and W. S. Kim. Dose modification of alemtuzumab in combination with dexamethasone, cytarabine, and cisplatin in patients with relapsed or refractory peripheral t-cell lymphoma: analysis of efficacy and toxicity. *Investigational new drugs*, 30(1):368–375, 2012.

[24] F. Guilhot, C. Chastang, M. Michallet, A. Guerci, J.-L. Harousseau, F. Maloisel, R. Bouabdallah, D. Guyotat, N. Cheron, F. Nicolini, et al. Interferon alfa-2b combined with cytarabine versus interferon alone in chronic myelogenous leukemia. *New England Journal of Medicine*, 337(4):223–229, 1997.

[25] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, H. Nordgren, et al. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research*, 66(21):10292–10301, 2006.

[26] E. Maubec, V. Chaudru, H. Mohamdi, F. Grange, J.-J. Patard, S. Dalle, B. Crickx, B. B.-d. Paillerets, F. Demenais, and M.-F. Avril. Characteristics of the coexistence of melanoma and renal cell carcinoma. *Cancer*, 116(24):5716–5724, 2010.

[27] D. J. Becker, S. Talwar, B. P. Levy, M. Thorn, J. Roitman, R. H. Blum, L. B. Harrison, and M. L. Grossbard. Impact of oncology drug shortages on patient therapy: Unplanned treatment changes. *Journal of Oncology Practice*, 2013.

A Appendix

A.1 Algorithms

Let the shortest distance matrix of graph G_{DDs}, the bipartite graph consisting of edges between chemotherapeutic drugs and cancers they treat, be D. We perform PCA on the set of row vectors of D, and project all vectors into the vector space defined by the principal components. We then discard all but the 20 principal components that contribute the highest variance, and then project all vectors back into the original vector space. We form a modified matrix D' with these vectors as row vectors. We then redraw G to form G', in which an edge between i and j exists if and only if $D'_{ij} \leq 1.5$.

A Bayesian clustering algorithm due to Vazquez [13] was then applied to the modified hypergraph. From G', we defined a hypergraph H, with hyperedges corresponding to the disease adjacency sets of the drugs in the database. Then define a to be the adjacency matrix of H; that is, $a_{ij} = 1$ if and only if vertex i belongs to edge j.

Let B denote the Beta function. We define:

$$B(p;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}; D(\pi;\gamma) = \frac{1}{B(\gamma)} \prod_{k=1}^{K} \pi_k^{\gamma_k - 1}$$

A similar limit applies for γ in this case.

We apply a Variational Bayes expectation-maximization (EM) algorithm. We minimize an upper bound F on the negative log likelihood of the data by performing a convergent algorithm that converges at the most likely cluster assignment probability matrix and adjacency probability matrix. Variational approximations for probabilities and parameters are first computed, followed by cluster probabilities, at every step of the convergent algorithm.

Let K be an initial upper bound for the appropriate number of clusters. Let θ_{kj} be the probability that a vertex in cluster k will belong to hyperedge j. Let p be a matrix of probabilities of cluster assignments. Let π denote the hidden frequency vector describing cluster sizes. This frequency vector is relevant because it allows us to define a prior $D(\pi;\gamma)$ that punishes uneven clusterings. Let γ denote a vector, indexed by group indices k, where γ_k refers to the sum of the probabilities of each node falling in cluster k. In all of our equations, $R(\theta), R(\pi)$ denote likelihood estimates for θ, π, respectively. Also, define

$$\langle A(\phi) \rangle = \int d\phi P(\phi|D) A(\phi),$$

for any function A on the parameters ϕ.

To ensure convergence at a global minimum of Kullback-Leibler (KL) divergence, we seed the vector π with a large number (1000) of random sets of probabilities, and the following algorithm is applied to each π, after which the parameters corresponding to the lowest KL divergence are chosen as the final parameters for our model.

Until F varies by less than a certain threshold ϵ, the following steps are iterated (we chose 10^{-6}):

$$m_{ik} = \langle \ln \pi_k \rangle + \sum_j a_{ij} \langle \ln \theta_{kj} \rangle + (1 - a_{ij}) \langle \ln(1 - \theta_{kj}) \rangle$$

$$p_{ik} = \frac{e^{m_{ik}}}{\sum_s e^{m_{is}}}$$

$$R(\theta) = \prod_{kj} B(\theta_{kj}; \alpha_{kj}, \beta_{kj})$$

$$\langle \ln(\theta_{kj}) \rangle = \psi(\alpha_{kj}) - \psi(\epsilon + \gamma_k)$$

$$\langle \ln(1 - \theta_{kj}) \rangle = \psi(\epsilon - \gamma_k - \alpha_{kj}) - \psi(\epsilon + \gamma_k)$$

$$\alpha_{kj} = \epsilon + \sum_{ij} p_{ik} a_{ij}$$

$$R(\pi) = D(\pi;\gamma)$$

$$\langle \ln(\pi_k) \rangle = \psi(\gamma_k) - \psi\left(\sum_k \gamma_k\right)$$

$$\gamma_k = \epsilon + \sum_i p_{ik}$$

$$F = \sum_{ik} p_{ik} \ln p_{ik} - \sum_{kj} \ln B(\alpha_{kj}, \beta_{kj}) - \ln B(\gamma)$$

Similar equations are defined for β.

After the algorithm finishes, the elements of p are the desired probabilities. Summing over all potential cluster assignments, we can then estimate the likelihood that a certain drug works on a certain disease.

To determine the quality of the clusters derived from the Bayesian algorithm, Fisher's exact test was applied in an enrichment analysis of treatment information. Clinical interpretations were assigned to each cluster, consisting of treatments shared among the diseases in the cluster. The frequency of the occurrence of these shared drug hyperedges in that cluster was then compared to the frequency in G_{DDs}, from which p-values were derived.

As the Bayesian algorithm calculates $\phi'_{kj} = \log \theta_{kj}$ and p_{ik}, we can determine the exact model-derived likelihood

$$P(i \in G_j | \phi) = \sum_k e^{\theta'_{kj}} p_{ik}, \tag{1}$$

where G_j represents the neighborhood of treatment j. Clusters resulting from the Bayesian hypergraph clustering algorithm were filtered by confidence, and only cluster assignments with confidences higher than 0.95 were retained.

The PCA-based topological abstraction algorithm was run using 20 principal components. After topological abstraction was complete, the Bayesian hypergraph clustering algorithm was applied to the resulting graph, excluding edges between pairs of drugs and pairs of diseases. Likelihoods were calculated using Equation (1).

A.2 Glossary

- 5-FU: 5-fluorouracil
- ALL: Acute lymphocytic leukemia
- AML: Acute myelogenous leukemia
- APL: Acute promyelocytic leukemia
- CLL: Chronic lymphocytic leukemia
- CML: Chronic myelogenous leukemia
- FL: Follicular lymphoma
- HCC: Hepatocellular carcinoma
- HIV NHL: Human immunodeficiency virus-related non-Hodgkin lymphoma
- HL: Hodgkin lymphoma
- H&N: Head and neck carcinoma
- MCL: Mantle cell lymphoma
- MDS: Myelodysplastic syndrome
- MPD: Myeloproliferative disorders
- MTOR: Mammalian target of rapamycin
- MZL: Marginal zone lymphoma
- NET: Neuroendocrine tumor
- NHL: Non-Hodgkin lymphoma
- NSCLC: Non-small cell lung cancer
- PCNSL: Primary central nervous system lymphoma
- R-CHOP: Rituximab, cyclophosphamide, hydroxydaunorubicin, Oncovin, prednisone
- SCLC: Small-cell lung cancer
- T-NHL: T-cell non-Hodgkin lymphoma

Toward a Cognitive Task Analysis for Biomedical Query Mediation

Gregory W. Hruby, MA[1], James J. Cimino, MD[1, 2], Vimla Patel[1,3] and Chunhua Weng, PhD[1]

1. Department of Biomedical Informatics, Columbia University, New York City;

2. NIH Clinical Center, Bethesda, Maryland;

3. New York Academy of Medicine, New York City

Abstract

In many institutions, data analysts use a Biomedical Query Mediation (BQM) process to facilitate data access for medical researchers. However, understanding of the BQM process is limited in the literature. To bridge this gap, we performed the initial steps of a cognitive task analysis using 31 BQM instances conducted between one analyst and 22 researchers in one academic department. We identified five top-level tasks, i.e., clarify research statement, explain clinical process, identify related data elements, locate EHR data element, and end BQM with either a database query or unmet, infeasible information needs, and 10 sub-tasks. We evaluated the BQM task model with seven data analysts from different clinical research institutions. Evaluators found all the tasks completely or semi-valid. This study contributes initial knowledge towards the development of a generalizable cognitive task representation for BQM.

Introduction

Helping researchers access "Big Data" in the electronic health record (EHR) is essential for both public health initiatives and comparative effectiveness research (CER) in many academic medical centers,[1,2] but remains a costly endeavor.[3] In reality, CER involves complex, ultra-granular information needs that necessitate assistance from data analysts to extract representative data from the EHR. To do this, the medical researcher's information need must be transferred to the data analyst, who may translate that information need into a precise and specific data query. The transfer of the information need from the medical researcher to the data analyst occurs through an iterative question-answering process. From this point on, we will refer to the transfer of the information need as the Biomedical Query Mediation (BQM) and a medical researcher may be any type of researcher seeking EHR data. During BQM, the data analyst may explain to the medical researcher relevant information of data restrictions and contextual data constraints, e.g., laboratory results may be more accurate than ICD-9 codes for identifying diabetes patients. Such information may guide medical researchers to reconsider and revise their queries.

Analogous to reference interview or interactive information retrieval in the field of Library and Information Science, the success of BQM depends on the effectiveness of the iterative negotiations between the data analyst and the medical researcher.[4-6] Reference interview elicits a clear and well-defined statement from the patron detailing the information need. However, literature provides neither rich insights into opaque BQM processes nor differences between locating data elements in the massive EHR information space and searching for books in libraries.[7]

Prior studies have shown that modeling interactive retrieval processes can lead to better designs of information retrieval systems.[5,8,9] As these studies suggest that and obscure processes, such as BQM, are observed by the few performers, and the resulting understanding of the process is limited. We believe that modeling the knowledge of the series of tasks performed by data analysts during BQM is important for providing standard, user-centered support for data analysts and medical researchers.

We previously reported a content analysis of the BQM between a data analyst and medical researchers.[10,11] As a natural extension to that study, this paper presents a cognitive task analysis of BQM to illustrate the BQM tasks and knowledge required to perform each task.[9,12] The purpose of a cognitive task analysis is twofold, first to outline the specific tasks used to accomplish a goal and second, to detail both the controlled and automated knowledge needed to perform each task identified.[13] A cognitive task analysis contains five core steps, i.e., collect preliminary information, identify task knowledge representations, apply focused knowledge elicitation methods, analyze and verify data acquired, and format the cognitive task analysis results for the intended application.[12] This study focuses on the first two steps. We modeled task activities and sequences, and knowledge needed to perform each task. Our task model underwent a face and content validation by external reviewers. Columbia University Medical Center Institutional Review Board approved this study. The rest of this paper first reports the methods and results and then discusses the implications of these findings for enhancing the BQM process.

Methods

Data Collection: Between July 2011 and January 2012, 31 discussions between one data analyst and 22 medical researchers were recorded and transcribed.

Data Analysis: Our analysis focused on the tasks occurring during BQM to accomplish transfer of the medical researcher's information needs to the data analyst. We extended our previous work's description of BQM constructs to seed BQM task identification.[11] The seeding content used was (1) the research question, (2) the clinical process, and (3) EHR data elements locations. Through a random selection of BQM transcripts and e-mails we initially identified tasks related to the seeded content and extrapolated sub-tasks related for each of these tasks. Through an iterative task and sub-task identification and refinement process, a final task and sub-task list emerged. Next, we ordered the task list temporarily. As prescribed by cognitive task analysis, we elaborate on the individual task attributes by identifying the task goals and knowledge required to complete each task. Additionally, we constructed a BQM knowledge representation in the form of hierarchical task complexity.

Face and Content Validity Evaluation: We presented our BQM task model to seven external data analysts from medical centers that are known to engage in BQM between data analysts and medical researchers (Northwestern University and Columbia University). We asked the data analysts to fill out a 14-item questionnaire for the derived BQM task list. Two items asked the data analyst to provide their experience with and frequency performing BQM. We used two items to assess face validity on a scale from 1-10 for the dimensions of representativeness of and usefulness for BQM. If the median score was greater than or equal to 7 for representativeness and usefulness, we considered the BQM task list to have face validity. Content validity is a metric determining whether a representation is capable of performing its intended task. Ten items were used to measure content validity; each data analyst judged the 10 sub-tasks as essential, useful, or non-useful. Inter-rater agreement in the form of content validity ratio was applied to assess content validity. Task content validity were achieved if the content validity ratio reaches the minimum critical value of 0.620.[14] Tasks were deemed semi-valid if at least half of the evaluators rated the task as essential.

Results

BQM task complexity: **Figure 1** presents the hierarchical complexity representation of the BQM tasks. We identified five tasks, i.e., define research statement, illustrate clinical process, identify related data elements, locate EHR data elements, and end mediation, with ten corresponding sub-tasks. A typical BQM contains iterative topic switching; therefore, the BQM process is not a linear progression among these tasks. Descending the pyramid, each task is broken down to relatively simpler though still quite complicated tasks. This hierarchy also highlights the iterative process between both the clinical process with EHR data element location and other data elements with EHR data element location. As the medical researcher becomes aware of what they do not know about what they think they may know as it pertains to available data elements within the EHR, one of two things may happen: i.e., the BQM may stop or the medical researcher may revise the research statement to accommodate the new knowledge gained through the BQM.

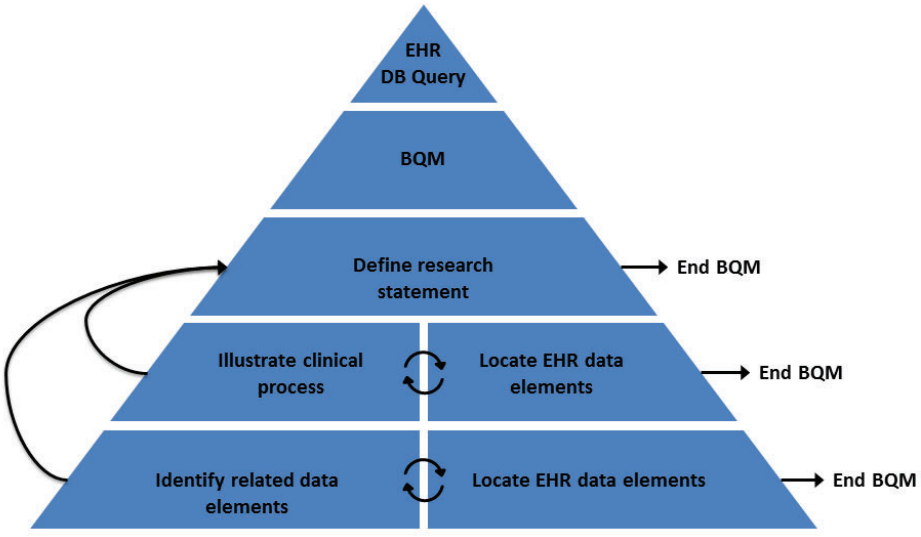


Figure 1: The complexity hierarchy and task flow for BQM

BQM task process and dimensions

Table 1: BQM tasks and activities performed by the data analyst

Task	Sub-task	Goal	Knowledge Required	Example
1. Define research statement	1.1 Elicit the clinical research scenario	To introduce core data elements of the information need	Study types	*What is the research question?*
	1.2 Understand the design of the proposed research	To establish the relationships among data elements	Study types	*Are you looking at pre-treatment factors that affect the outcome measure?*
2. Illustrate clinical process	2.1 Elicit the clinical progression related to the information need	To establish the temporal order of abstract data elements	Medical domain knowledge	*Patients with disease **x** that undergo treatment **y**, can you describe the diagnosis, treatment and follow-up timeline?*
	2.2 Gather specific details and data representations of the ordered abstract data elements	To establish EHR data definitions for abstract data elements	Medical domain knowledge	*Do all doctors refer to treatment X as x? What billing codes/image studies/lab tests are used for that type of visit?*
	2.3 Create list of unknown data elements	To provide inputs for task 3	Heuristics	*What is the data element X? Please describe.*
	2.4 Understand how to calculate derived variables from EHR data elements	To provide calculation parameters for derived variables	Heuristics	*The Duke University risk score takes into account variables x and y using this formula, x/y + 5.*
3. Identify related data elements	3.1 Elicit relevant abstract data elements not represented in the clinical process	To establish static variables required for the study	Medical domain knowledge	*What demographic information do you need? Any specific comorbidities?*
4. Locate EHR data elements	4.1 Show or request to see the location of the EHR data element	To establish location of data element within the data model of the EHR	EHR data model; EHR graphical user interface	*I'm unfamiliar with the data element X, where is it recorded in the EHR?*
	4.2 Describe availability and consistency of data elements	To educate the medical researcher on data quality, accessibility and reliability	EHR data model; Data quality, accessibility, and reliability	*Data element X is not collected in the EHR; Data element Y is available sporadically from patient to patient.*
5. End mediation	5.1 Inform the medical researcher whether or not the information need can be satisfied	To allow the medical researcher to reformulate their information need or end the BQM	EHR data model; Data quality, accessibility, and reliability	*That data element is contained in a scanned image and can't be extracted from the EHR.*

Evaluator Characteristics: Of the seven evaluators, 29% (2/7), 43% (3/7), and 29% (2/7) have been facilitating data access for >10, 3-5, and 1-2 years, respectively; 29% (2/7), 43% (3/7), 14% (1/7) and 14% (1/7) facilitate >10, 5-10, 3-5, and 1-2 BQM per month.

Face and content evaluation results: **Table 2** shows the score distribution and content validity ratio of the 10 items for the 10 sub-tasks from Table 1.

Table 2: Task Content Validation Results

Sub-task	Essential (%)	Useful (%)	Non-Useful (%)	Content Validity Ratio
1.1	71	29	0	0.43
1.2	71	29	0	0.43
2.1	57	43	0	0.14
2.2	100	0	0	**1**
2.3	71	29	0	0.43
2.4	86	14	0	**0.71**
3.1	71	14	14	0.43
4.1	71	14	14	0.43
4.2	100	0	0	**1**
5.1	100	0	0	**1**

The first item addressing face validity, on a scale from 1 (not at all)-10 (completely), rating 'to what extent does the task model simulate BQM', the median score was 8 (7-10). The second item addressing face validity, on a scale from 1 (not useful)-10 (very useful), rating 'how useful is this representation for novice data analysts conducting BQM' the median score was 8 (6-10).

Discussion

We identified five tasks and 10 sub-tasks used to elicit the medical researcher's information needs to the data analyst. Additionally, the BQM tasks were categorized into a complexity hierarchy. This representation serves as an initial theoretical framework for BQM. A closer look at similar frameworks used for interactive information retrieval and mediated searching shows some similarities. Specifically, the ASK hypothesis and Berrypicking model share similar task progressions to arrive at a clearer definition of an information seeker's need.[15,16] Additionally, Spink's proposed theoretical framework details seven levels for mediated searching: Problem solving process, information seeking episodes, uncertainty, cognitive styles, interactive search sessions, successive search behavior and sets of situated actions.[17] The proposed BQM does not necessarily overlap with these levels, but provides further depth for the interactive search session, or the dialogue between and a user and a system, in this case the channel to the system, the data analyst. We postulate the level of granularity present in EHR databases contributes significantly to the complexity of BQM. Unlike document databases, EHR database complexity is present in both the breadth of data elements and the features used to describe those data elements. Our theory resonates with other interactive information retrieval experts. Ford et al states, "The deeper and more structured are the knowledge representation formalism adopted, the more difficult it is to develop systems able to accommodate wide-ranging subject content.[18]" System development in this context may be used as a surrogate for the model of the proposed work flow the system is attempting to improve. As such, it can be inferred that increased information breadth and knowledge representation granularity will increase the complexity of the interaction with an information retrieval system.

Complexity hierarchy: Our top-level concept explains the ultimate goal of BQM, querying the EHR database. However, the concepts below it break this high-level concept into simpler atomic concepts. BQM, a critical, but yet complex component of EHR database queries is broken further down into the critical components, first the intent of the information need, what is the research the medical researcher wants to conduct. This is then followed by more granular components of assigning clinical data elements to a clinical time line and explicitly defining abstract elements into representations within the EHR. Additionally, other data characteristics that may not be directly tied to a clinical time are explored and the EHR representation is also defined. This model suggests two feedback loops and several potential BQM stopping points. As the data analyst becomes aware of the medical researcher's information needs the data analyst has a clearer understanding of EHR information space's ability to contain a representative dataset. Similarly, as the medical researcher's awareness of the unknown and what is thought to be known may affect the initial intent thereby augmenting the research statement to accommodate what EHR data are suitable to represent an information need. Of note, knowledge required to perform all tasks come from different sources. These sources aid the data analyst to locate and map data elements to the internal data repository for the medical researcher. Both medical researchers and data analysts can contribute medical knowledge; that is why the conversation between the two is crucial.

BQM task process model: Of particular interest, task 1, define research statement, mirrors a similar process to the reference interview to understand the context for which the informant seeker is working from. Understanding the intent of the researcher provides a scaffolding of core information elements and the relationships of the information elements.[4,19] Likewise, the initial task of BQM enables the data analyst to develop an internal information model similar to the one being used by the medical researcher MR. This mental model is a semantic relationship of the key medical data elements for which finer details related to those elements can be explored.

Tasks 2, 3, and 4 are used to facilitate the focus of the information need. It has been shown that the focusing of the information seeker's need provides an increases in precision of the results.[7] While this study did not assess the precision of resulting datasets, tasks 2-4 support the notion of an iterative refinement and understanding of the medical researcher's information need. Each task builds from the initial task, and the subsequently these task may augment the initial task or end the BQM. The final task represents the data analyst understanding of the medical researcher's information need and whether or not that need can be met by EHR data. This task can either end the mediation with no results or move the mediation into a formal EHR database query.

Face and content validation: Our expert evaluators deemed the preliminary process model to have face validity. The ratings suggest our initial representation of BQM is both representative of and useful for BQM. Additionally, 40% (4/10) of the sub-tasks, sub-tasks 2.2, 2.4, 4.2, and 5.1, were judged to have content validity. All tasks were

judged either semi-valid or valid. Given these positive results, we believe this initial knowledge representation of BQM is acceptable to continue with our cognitive task analysis by applying focused knowledge elicitation methods with data analysts representing diverse approaches to BQM.

Limitations: Our study contains several limitations. This representation is based on a specific data set covering the BQM process for just one data analyst and one medical research domain, urologic oncology. Our findings may not be representative of information needs from other medical domains, nor may it be inclusive of other expert data analyst tasks used to transfer a medical researcher's information need to them. Additionally, the EHR data being accessed was represented by an integrated research data repository, which contained highly granular data. The type of EHR information source may affect the BQM process. Regardless of these limitations, the initial components of the cognitive task analysis will allow us to move forward with informed semi-structure interviews of other data analyst experts with the sole purpose of extracting additional knowledge of BQM.

Conclusions

This study contributes preliminary knowledge of BQM task sequence and a task complexity knowledge representation. This knowledge can guide future work on cognitive task analysis for acquisition of additional information from a diverse group of data analysts on the tasks used to accomplish a generalizable BQM.

Acknowledgments

This research was funded under NLM grant **R01LM009886, R01LM010815, and 5T15LM007079,** and CTSA award **UL1 TR000040**. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIH. The authors would like to thank Luke Rasmussen (Northwestern University) for helping with recruiting evaluators for our study.

References

1. D'Avolio LW, Farwell WR, Fiore LD. Comparative effectiveness research and medical informatics. *The American Journal of Medicine.* 2010;123(12):e32-e37.
2. Hoffman S, Podgurski A. Big Bad Data: Law, Public Health, and Biomedical Databases. *Public Health, and Biomedical Databases (October 30, 2012). Journal of Law, Medicine and Ethics, Forthcoming.* 2012.
3. Rein A. Finding Value in Volume: An Exploration of Data Access and Quality Challenges. *AcademyHealth: Briefs and Reports.* 2012:9.
4. Taylor RS. *QUESTION-NEGOTIATION AN INFORMATION-SEEKING IN LIBRARIES*: DTIC Document;1967.
5. Kuhlthau CC. Inside the search process: Information seeking from the user's perspective. *JASIS.* 1991;42(5):361-371.
6. Spink A. Study of interactive feedback during mediated information retrieval. *Journal of the American Society for Information Science.* 1997;48(5):382-394.
7. Vakkari P. Task-based information searching. *Annual review of information science and technology.* 2005;37(1):413-464.
8. Suchman L. Making work visible. *Commun. ACM.* 1995;38(9):56-ff.
9. Weir CR, Nebeker JJ, Hicken BL, Campo R, Drews F, LeBar B. A cognitive task analysis of information management strategies in a computerized provider order entry environment. *Journal of the American Medical Informatics Association.* 2007;14(1):65-75.
10. Hruby GW W, A, Weng C. Analysis of Query Negotiation between a Researcher and a Query Expert. Paper presented at: AMIA; 1780, 2012; Chicago.
11. Hruby GW BM, Cimino JJ, Gao J, Wilcox AB, Hirschberg J, Weng C. Characterization of the Biomedical Query Mediation Process. *AMIA Summits on Translational Science Proceedings*. San Francisco2013:5.
12. Schraagen JM, Chipman SF, Shalin VL. *Cognitive task analysis*: Lawrence Erlbaum; 2000.
13. Clark RE, Estes F. Cognitive task analysis for training. *International Journal of Educational Research.* 1996;25(5):403-417.
14. Wilson FR, Pan W, Schumsky DA. Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development.* 2012;45(3):197-210.
15. Belkin NJ, Oddy RN, Brooks HM. ASK for information retrieval: Part I. Background and theory. *Journal of documentation.* 1982;38(2):61-71.
16. Bates MJ. The design of browsing and berrypicking techniques for the online search interface. *Online Information Review.* 1989;13(5):407-424.
17. Spink A, Wilson TD, Ford N, Foster A, Ellis D. Information- seeking and mediated searching. Part 1. Theoretical framework and research design. *Journal of the American Society for Information Science and Technology.* 2002;53(9):695-703.
18. Ford N, Ford R. Towards a cognitive theory of information accessing: an empirical study. *Information Processing & Management.* 1993;29(5):569-585.
19. Ross CS, Nilsen K, Dewdney P. *Conducting the Reference Interview: A How-to-do-it Manual*: Neal-Schuman; 2002.

Cross-System Evaluation of Clinical Trial Search Engines

Silis Y. Jiang[1], BS, Chunhua Weng[1], PhD
[1]Department of Biomedical Informatics, Columbia University, New York, NY 10032

Abstract

Clinical trials are fundamental to the advancement of medicine but constantly face recruitment difficulties. Various clinical trial search engines have been designed to help health consumers identify trials for which they may be eligible. Unfortunately, knowledge of the usefulness and usability of their designs remains scarce. In this study, we used mixed methods, including time-motion analysis, think-aloud protocol, and survey, to evaluate five popular clinical trial search engines with 11 users. Differences in user preferences and time spent on each system were observed and correlated with user characteristics. In general, searching for applicable trials using these systems is a cognitively demanding task. Our results show that user perceptions of these systems are multifactorial. The survey indicated eTACTS being the generally preferred system, but this finding did not persist among all mixed methods. This study confirms the value of mixed-methods for a comprehensive system evaluation. Future system designers must be aware that different users groups expect different functionalities.

Introduction

Clinical trials are the gold standard for establishing the effectiveness or efficacy of new drugs and treatments. One of the challenges to successfully completing clinical trials is recruiting enough research participants[1]. An approach to increase the public's awareness of clinical trials is the publicly accessible clinical trial repository. Currently, several countries have such repositories[2]. In the United States in 2007, the FDA mandated that all new U.S.-based clinical trials be registered with the repository ClinicalTrials.gov[1].To date, ClinicalTrials.gov contains more than 153,260 clinical trials[3]. While ClinicalTrials.gov offers centralized access to these trials, searching for relevant trials remains difficult for the average user[2]. Often, the complex and technical language used to describe a trial and its eligibility criteria are difficult for a user to comprehend[4].

Anecdotally, the largest user group of ClinicalTrials.gov consists of those who seek to participate in a clinical trial. Besides using ClinicalTrials.gov, many users are also utilizing other ClinicalTrials.gov extension systems to search for trials. Many commercial clinical trial search engines have emerged, including Corengi.com[5], TrialX.com[6], eTACTS (http://is.gd/eTACTS)[7], Patientslikeme.com[8], and TrialReach[9]. Most are disease-specific, such as Dory/TrialX (cancer) or Corengi (diabetes mellitus type II). Others, such as Patientslikeme, allow users to search for clinical trials related to specific disease types.

While clinical trial repositories and search engines have existed for more than a decade now, there is little research on how users search for and process clinical trial information. In 2008, Atkinson conducted an expert usability analysis of various cancer-specific clinical trial search engines[4]. In 2010, Patel conducted a study to identify the common search queries of users searching for clinical trial information[10]. A key aspect of promoting the use of these search tools is developing refinements that increase user acceptance and adoption of these tools. Since increasing usability is a critical aspect of user acceptance[11], we felt that the lack of understanding of both users and systems in this environment could lead to inadequate designs. Motivated to bridge this important knowledge gap, we conducted this study of a few representative clinical trial search engines to understand how users search for clinical trials and interpret the information presented to them by these search engines. The Institutional Review Board (IRB) at Columbia University Medical Center approved this study.

Materials and Methods

1. Selected Clinical Trial Search Engines

We compared eTACTS (the system developed at Weng's lab at Columbia) with four other representative clinical trial search engine sites: ClinicalTrials.gov, Corengi, Dory offered by TrialX, and PatientsLikeMe. **Table 1** summarizes the key functionalities of each. ClinicalTrials.gov requires a user to complete a simple or advanced form to specify search queries for clinical trials. Corengi requires a user to create a personal profile to allow automated matching of trials to user characteristics. PatientsLikeMe presents to users pre-selected demographic questions to help them filter clinical trials. Dory, an online intelligent agent provided by TrialX, enables interactive search between a user and a human customer service agent. eTACTS uses a dynamic cloud of pre-mined tags to allow a

user to filter clinical trial search results from ClinicalTrials.gov.

2. Study Participants

We recruited 11 participants. We tried to increase the heterogeneity of our sample by recruiting physicians, research coordinators, and medical novices. Our study participants also had varied levels of computer skills. A complete summary of the participants can be found in **Table 2**. Some participants were initially contacted by a third-party research coordinator and referred to our research team. For these participants, our research team then contacted the potential participant for an evaluation session. Other participants were recruited directly by the research team through an email recruitment announcement within our academic department. Internal recruitment was limited to participants who had not previously used eTACTS .

Table 1. Summary of The Representative Clinical Trial Search Engines (N=5)

System	Key Functionalities
ClinicalTrials.gov	Offers both simple and advanced searches using string-based free-text search
Corengi	Matches patients up to clinical trials based on user-provided profile information
Dory/TrialX	Provides summaries and contact information to users based on question and answer sessions
eTACTS	Provides interactive tag cloud to allow users to select clinical terms to filter clinical trials
PatientsLikeMe	Provides a set of pre-formed search queries to help filter clinical trials

Table 2. Participant Diversity (N=11)

Diversity Dimension	Proportion	Percentage of Sample (n=11)
Male	10	90.9%
Clinicians (MD)	3	27.3%
Database Administrators	3	27.3%
Graduate Students	3	27.3%
Clinical Research Coordinators	2	18.2%
Experienced Users	6	54.5%

3. Scenario-based mixed-methods evaluations

We obtained informed consent to participate in this study from all the participants before the study began. Participants then completed the one-hour long evaluation session and a short debriefing session followed to answer any remaining questions. To evaluate and compare these systems, we asked participants to search for clinical trials and determine whether a mock patient would be eligible for a specific trial. Some tasks required functions unique to a particular search engine (**Table 1**), while others used functions common to all search engines.

We used a mixed-methods evaluation design in order to measure several levels of behavior and usage patterns. In the absence of usage logs, we used time-motion analysis to capture system usage time. We complemented the quantitative time-motion analysis with two qualitative methods to capture user preferences, opinions about the systems, and any other unexpected findings. To allow users to explore each system, we devised a set of tasks for each participant to accomplish. To help give participants a reference, our team supplied the participants with mock-patients diagnosed with diabetes mellitus type II. We chose to use this disease for our mock patient due to its high prevalence in the clinical trial repository and in the population. We created mock patients by combining several eligibility criteria typical of trials studying this disease. The following is an example of a mock patient:

> *You are a 40-year-old Caucasian male (born in March, 1973) with type II diabetes mellitus diagnosed in June 2005. You take the anti-diabetic drug metformin. You have a hemoglobin A1c value of 7.3. You weigh 220 pounds, are 5 feet 10 inches tall, and have a body mass index (BMI) of 32. You do not smoke and you drink socially. You lead a sedentary life and live with your wife. You have mild hypertension, which is medically controlled with an anti-hypertensive diuretic medication. You have no other significant co-morbidities. You see a primary care physician regularly and have commercial health insurance.*

For every participant, we used ClinicalTrials.gov as the baseline search engine to train the participant how to search for clinical trials and how to understand eligibility search criteria. In order to simulate the real-world environments, the participants were all self-trained. After completing the baseline tasks, we selected participants to systematically complete tasks for each system, with a random sequence of systems. In order to collect qualitative

data, we asked participants to verbalize their thoughts while completing the tasks. We instructed users to speak out loud any thoughts, expectations, and surprises about the system. We recorded their voices using a separate audio recorder. To capture the task completion time of each user, we used the TimeCaT tool.[12] TimeCaT allows an investigator to time a participant completing various tasks. Tasks can be defined either *a priori* or *ad hoc*. To define our tasks *a priori*, we completed the scenarios using each system and created an activity diagram for each system. Each step represented in the activity diagram was then converted into a task to be recorded and timed during the time-motion analysis. In this study, we recorded the amount of time required by participants to enter information into the system, to execute the key system functions, and to determine whether the mock patient was eligible for the trials returned by the search engine. **Table 3** lists the tasks observed in this study.

Table 3. Time-Motion Analysis Task List

Task	Description	Category
Typing Information	User enters initial search query information (i.e. diabetes mellitus type II) into the clinical trial search engine.	Preparatory
Answer Questions	User responds to a set of iterative questions (only used for Dory)	Preparatory
Refine Tag Cloud	User reviews and selects tag cloud options (only used for eTACTS)	Preparatory
Entering Profile	User enters information required to establish medical profile (only used for Corengi and PatientsLikeMe)	Preparatory
System Interactions	Time required by clinical trial search engine to process information and return response, or time spent by user on navigating the interface is also included.	Interaction
Result Review	User reviews the returned list of clinical trials. Participants may determine mock patient eligibility at this stage.	Review
Trial Review	User reviews a single clinical trial to determine whether mock patient would qualify for the study.	Review

Additionally, we used a modification of a survey by Zheng et al. to measure the user perceptions of each system[13].This survey is based on the Unified Theory of Acceptance and Use of Technology[13]. In order to encourage participants to compare the systems with each other, the majority of the survey questions asked participants to rank the systems by a set of system features preferred by users. Participants took the survey immediately after completing the scenario-based task evaluation portion of the study.

4. Statistical Analysis

In order to test for differences between groups, we used a battery of non-parametric test. To calculate the time usage difference between various user groups (i.e. clinicians and non-clinicians), we used the Mann-Whitney U test[14]. To analyze difference between time spent in each task category, we used the Kruskal Wallis test[14] and the Bonferroni correction for post-hoc analysis[15]. We repeated the same statistical method to calculate the differences between survey score results.

1. Time-Motion Analysis

As shown in **Table 4**, after averaging all user time-motion data for each system, we found that the test users required the least time to interact with TrialX (avg=7.52 mins) and the most time to interact with eTACTS (avg=9.91).

When we separated user groups by degree of medical knowledge, in this case physicians versus non-physicians, physicians were not significantly faster than non-physicians, except in the case of eTACTS ($p=0.048$). As shown in Column D in **Table 4**, we found that physicians spent significantly less time completing the required tasks using eTACTS when compared to their counterparts.

We also compared those with experience searching for clinical trials with those having no such experience, as shown in Column G in **Table 4**. eTACTS and PatientsLikeMe lead to large time usage differences between these two user groups. We observed the biggest difference between experienced and novice user interaction time with the eTACTS system ($\Delta=2.54$).

When we aggregated all system usage times together, we found no statistical difference between the two sets of user groups (physicians versus non-physicians, experienced user vs. non- user).

Table 4. Average time spent by user groups per system (A: average time spent; B: average physician time spent; C: average non-physician time spent; D: difference between B and C; E: average experienced user time spent; F: average novice user time spent; G: difference between E and F. All measures are in minutes; 🡻 indicates the ranking column.

System	A (🡻)	B	C	D=C-B	p-value (D)	E	F	G=E-F	p-value (G)
Dory/TrialX.com	7.52	7.27	7.62	0.35	0.776	8.02	6.93	1.09	0.429
Corengi	8.19	8.05	8.22	0.22	1.00	7.86	8.51	-0.65	0.310
ClinicalTrials.gov	8.44	7.74	8.71	0.97	0.63	7.30	9.82	-2.52	0.247
PatientsLikeMe	9.78	7.99	10.45	2.46	0.776	8.22	11.65	-3.43	0.126
eTACTS	9.91	5.33	11.62	**6.29**	*** 0.048**	11.06	8.52	2.54	0.792
Average	**8.77**	**7.28**	**9.32**	**2.06**	**0.259**	**8.49**	**8.63**	**0.14**	**0.366**

Table 5. Average time spent per user per task group by system (all measures are in minutes;🡻 indicates the ranking column).

System	Interaction	Preparatory	Review	Other[#]	Total (🡻)
Dory/TrialX.com	0.39	2.83	3.97	0.33	7.52
Corengi	1.23	2.84	3.81	0.31	8.19
ClinicalTrials.gov	0.42	0.39	6.79	0.84	8.44
PatientsLikeMe	2.42	1.22	5.46	0.68	9.78
eTACTS	0.45	2.29	6.24	0.93	9.91
Average	**0.98**	**1.91**	**5.25**	**0.62**	

[#]Other represents tasks not originally intended to be recorded, such as soliciting for help or asking for clarification.

We found that PatientsLikeMe required more time devoted to site navigation than other systems. The Kruskal Wallis test was significant for differences in time required for interaction among different systems with post-hoc comparisons indicating that PatientsLikeMe and Corengi are significantly more time consuming than ClinicalTrials.gov ($p = 0.022$ and $p = 0.001$ respectively, **Table 5 Column 1**). Furthermore, Corengi was significantly more time intensive than eTACTS ($p = 0.005$, **Table 5 Column 1**).

Simple searches on ClinicalTrials.gov required the least overheard commitment from users before arriving at results, but this difference was not statistically significant ($p = 0.158$, **Table 5 Column 2**).

When breaking down the time spent by users per task, we can easily identify reviewing eligibility status as the most time-consuming task. ClinicalTrials.gov and eTACTS require the most time for users to determine their eligibility for each trial. However, the time spent on each system to review eligibility status was not significantly different ($p = 0.492$, **Table 5 Column 3**).

2. Think-aloud Protocol

Our audio recordings revealed a number of observations about the systems and clinical trials in general. One of the chief complaints that participants voiced was the ambiguity and/or complexity of a clinical trial's eligibility criteria. For instance, one clinical trial described two apparently contradicting criteria.

"*What does this criterion mean? How can you simultaneously have never taken a drug and have taken it for more than a month? There is an 'or' in this criterion! That was very unclear from a quick glance*." – Participant 1.

Some participants had difficulty understanding that the criteria called for participants having never taken medications before or having taken a stable dosage for more than 4 weeks. For these participants, the meaning of the criteria was lost in the long and highly specific wording.

With regard to specific systems, the voice recording demonstrates that personal preferences play a role in a participant's perception. For the Corengi system, two of the participants felt that completing the medical profile form online was a simple but time-consuming process. This reduced the value of using a medical profile to filter clinical trials for participants. Six participants felt that the Dory interface by TrialX was easy to use; however, an equal number felt that the information presented by Dory was inadequate for assessing clinical trial eligibility. This lack of information was dissatisfying for some participants:

"[*Dory] doesn't give me a lot of information. […] I would really need more information in order to determine whether the patient qualifies or not. […] This is really for – It's not for researchers to use… for people to find out what's out there*." – Participant 6.

Additionally, the rigid diagnosis vocabulary and strict diagnosis autocomplete function used by the Dory to identify the user's condition caused some participants to conduct searches that returned no results. For the eTACTS system, participants were split in their perceptions toward the tag cloud feature. While some found this feature useful for quickly filtering the possible clinical trials, others found it confusing:

"*The first few tags were easy to find, but afterwards, I became more unsure of which tags applied to me.*" – Participant 3.

One physician was unable to intuitively use the tag cloud feature, while the remaining clinicians found it difficult to understand some of the tags (e.g. "blood pressure") in the context of clinical trials. Participants in the study often found the PatientsLikeMe system very polished. Unlike the autocomplete function in Dory, all participants who commented on the system performance were impressed with the speed of the autocomplete function on PatientsLikeMe. Also, one participant noticed that the trials recommended by PatientsLikeMe differed from those recommended by ClinicalTrials.gov.

3. Survey

The post-scenario-based evaluation survey also presented some interesting results, which can be found in Table 6. Overall, eTACTS was consistently the favored search engine system. It received the best rating in four out of the five aspects surveyed. PatientsLikeMe was deemed to provide the most guided search. We found no statistical differences when comparing the various engines based on each of the survey questions. Since none of the Kruskal Wallis tests were significant, we did not use a post-hoc test.

Table 6. Average rating for search engines, which are ordered from left to right by ease of use using a 5-scale Likert survey (1: most preferred; 5: least preferred, A = eTACTS, B = Dory/TrialX, C = ClinicalTrials.gov, D = PatientsLikeMe, E = Corengi)

Aspect	A	B	C	D	E	P-value
Ease of entering information	2.42	3.58	2.92	3.00	4.00	0.172
Provided most search guidance	3.00	2.75	3.92	2.50	3.45	0.166
Ease of site navigation	1.75	2.83	2.17	3.17	3.36	0.173
Ease of use with no prior	2.08	2.75	2.92	3.33	3.18	0.351
Overall ease of use	*** 2.17**	**2.50**	**2.67**	**2.75**	**3.18**	**0.665**

Discussion

In our study, we applied mixed methods to evaluate five clinical trial search engines. Our results identified two important implications for designing clinical trial search engines. While eTACTS scored well on the survey, there was much variation in time spent using eTACTS in both of our user group comparisons. We found that eTACTS was not unique in this aspect; others such as PatientsLikeMe also required varied time among users. In the aggregate, eTACTS was one of the most time-consuming search engines. There are at least two possible explanations for this finding.

The first explanation is that time required for a user to find trials is not necessarily related to that user's satisfaction with the system. This suggests that other factors, such as those we measured in the survey, may also influence user satisfaction. Research into the usability of other information retrieval systems (IR) also found that aspects such as clarity of the interface design and ease of learning the interface impact the usability and user experience of the system[16]. In our think aloud protocol, we found that systems that required more attention and time spent navigating were often not favored and garnered more complaints. The findings in this study reinforce conclusions from the literature that developers should consider multiple user measures when evaluating the usability of search engine interface designs.

The second explanation is that individual variability influences the usability of these systems. A recent focus in information science has been the impact of users' cognitive processes and usability on web searching behavior[17]. In the case of eTACTS, we found that participants who favored the system the most were those who had extensive medical knowledge and an understanding of how the tag cloud feature worked. Individual variability was a factor in other systems as well. Kinley et al. found that individual cognitive styles (imager, verbalizer, etc.) could impact

which of the three information-processing approaches (scanning, reading, and mixed) was favored by an individual[18]. In our own study, participants daunted by the complexity of clinical trial eligibility criteria found Dory/TrialX.com appealing; while those who wanted to see the exact criteria found it frustrating. Kim et al. identified problem-solving style as another contributor to how users search for information[18]. Future work in this field should focus on identifying user groups for these systems and then identifying each group's design and functionality preferences. These types of findings would give insights to clinical trial search engine design while addressing multiple factors of technology acceptance and use.

One limitation of our study lies in our sample. It is both small (n = 11) and selective. Our participants all had a bachelor's degree (or higher) and are familiar with using Internet to seek health information; therefore, their behaviors and cognitive styles may not reflect that of the greater population, especially those who are older and are not familiar with using the Internet for information seeking. In spite of these potential limitations, we believe the types of participants we sampled are those who would most often use this type of resources. The main purpose of this study was to demonstrate the potential for greater emphasis of user design based on multiple perspectives. Further studies are needed to test how the results of our comparisons of clinicians and experienced trial searchers and their respective novice counterparts may generalize to the general population.

Conclusions

We presented a mixed-method approach to understand user interactions with five representative clinical trial search engines. We found that (1) user groups exhibit different behaviors when searching for clinical trials; (2) clinicians preferred certain system features to others; and (3) individual cognitive styles and characteristics affect user behaviors and usage patterns. Our study contributed empirical evidence that differences among users can influence user search behaviors and hence should be considered during the design of clinical trial search engines.

Acknowledgments

The research described was supported by grants **R01LM009886**, **R01LM010815**, and **T15LM007079** from the National Library of Medicine, and grant **UL1 TR000040** from the National Center for Advancing Translational Sciences (NCATS).

References

1 Institute of Medicine (US). *Envisioning a Transformed Clinical Trials Enterprise in the United States: Establishing An Agenda for 2020: Workshop Summary*. Washington (DC): : National Academies Press (US) 2012.

2 Dear R, Barratt A, Askie L, *et al.* Adding value to clinical trial registries: insights from Australian Cancer Trials Online, a website for consumers. *Clinical trials (London, England)* 2011;**8**:70–6. doi:10.1177/1740774510392392

3 *clinicaltrials.gov*. http://www.clinicaltrials.gov (accessed 29 Sep2013).

4 Atkinson NL, Saperstein SL, Massett HA, *et al.* Using the Internet to search for cancer clinical trials: A comparative audit of clinical trial search tools. *Contemporary Clinical Trials* 2008;**29**:555–64. doi:10.1016/j.cct.2008.01.007

5 *Corengi.com*. https://www.corengi.com (accessed 29 Sep2013).

6 *Clinicl Trials Search Service -- Powered by Ask Dory*. http://dory.trialx.com/ask/ (accessed 29 Sep2013).

7 Miotto R, Jiang S, Weng C. eTACTS: A Method for Dynamically Filtering Clinical Trial Search Results. *Journal of biomedical informatics* 2013;:1–8. doi:10.1016/j.jbi.2013.07.014

8 *patientslikeme.com*. http://www.patientslikeme.com (accessed 29 Sep2013).

9 *trialreach.com*. http://trialreach.com (accessed 9 Jan2014).

10 Patel CO, Garg V, Khan SA. What do patients search for when seeking clinical trial information online? *AMIA*

Annu Symp Proc 2010;**2010**:597–601.

11 Beyer H. *Contextual Design: Defining Customer-Centered Systems*. 1st ed. San Francisco: : Morgan Kaufmann 1997.

12 Lopetegui M, Yen P-Y, Lai AM, *et al.* Time Capture Tool (TimeCaT): development of a comprehensive application to support data capture for Time Motion Studies. *AMIA Annu Symp Proc* 2012;**2012**:596–605.

13 Zheng K, Padman R, Johnson MP, *et al.* An interface-driven analysis of user interactions with an electronic health records system. *J Am Med Inform Assoc* 2009;**16**:228–37.

14 Choosing the Correct Statistical Test in SAS, Stata and SPSS. ats.ucla.edu. http://www.ats.ucla.edu/stat/mult_pkg/whatstat/ (accessed 9 Jan2014).

15 Weisstein EW. Bonferroni Correction -- from Wolfram MathWorld. *mathworldwolframcom*

16 Zhang Y. Searching for specific health-related information in MedlinePlus: Behavioral patterns and User Experience. *J Am Soc Inf Sci Tec* 2013;:n/a–n/a. doi:10.1002/asi.22957

17 Kinley K, Tjondronegoro D, Partridge H. Web searching interaction model based on user cognitive styles. 2010;:340–3.

18 Kim KS, Allen B. Cognitive and task influences on Web searching behavior. *J Am Soc Inf Sci Tec* 2002;**53**:109–19.

Application of HL7/LOINC Document Ontology to a University-Affiliated Integrated Health System Research Clinical Data Repository

Yan Wang, MS[1], Serguei Pakhomov, PhD[1, 2], Justin L. Dale, BA[3], Elizabeth S. Chen, PhD[5, 6, 7], Genevieve B. Melton, MD, MA[1, 4]
[1]Institute for Health Informatics, [2]College of Pharmacy,[3] Academic Health Center - Information Systems, [4]Department of Surgery, University of Minnesota, Minneapolis, MN; [5]Center for Clinical & Translational Science, [6]Department of Medicine, [7]Department of Computer Science, University of Vermont, Burlington, VT

Abstract
Fairview Health Services is an affiliated integrated health system partnering with the University of Minnesota to establish a secure research-oriented clinical data repository that includes large numbers of clinical documents. Standardization of clinical document names and associated attributes is essential for their exchange and secondary use. The HL7/LOINC Document Ontology (DO) was developed to provide a standard representation of clinical document attributes with a multi-axis structure. In this study, we evaluated the adequacy of DO to represent documents in the clinical data repository from legacy and current EHR systems across community and academic practice sites. The results indicate that a large portion of repository data items can be mapped to the current DO ontology but that document attributes do not always link consistently with DO axes and additional values for certain axes, particularly "Setting" and "Role" are needed for better coverage. To achieve a more comprehensive representation of clinical documents, more effort on algorithms, DO value sets, and data governance over clinical document attributes is needed.

Introduction
Electronic Health Record (EHR) systems and electronic clinical documentation allow for functions such as decision support, quality assurance and clinical research. Clinical data warehouses and other infrastructures incorporating multiple data sources can enable researchers to perform cohort identification, hypothesis generation, retrospective analyses and other research functions[1]. With the rapid proliferation of locally customized documents in clinical care, there is a growing need for standard representations to enable their storage, navigation, retrieval and use.

The HL7/LOINC Document Ontology (DO) is an ontology for standardizing clinical documents with terms in a hierarchical structure to support exchange and reuse of clinical documents across institutions and different systems[2]. It is composed of five axes: *Kind of Document (KOD)* (e.g., Letter, Report, Note), *Type of Service* (*TOS*) (e.g., Consultation, Procedure, Evaluation and Management), *Setting* (e.g., Intensive Care Unit, Birthing Center*), Subject Matter Domain* (*SMD*) (e.g., Urology, Cardiovascular Disease) and *Role* (e.g., Physician, Registered Nurse, Technician). Each axis has a set of restricted values. In addition, many precoordinated value sets for DO map to LOINC codes.

In building the University of Minnesota clinical data repository as an infrastructure primarily for clinical and translational researchers in collaboration with University of Minnesota-affiliated Fairview Health Services, both legacy and current EHR system data are incorporated. The goal of this study was to represent clinical documents in a manner to facilitate their reuse by researchers for clinical studies and for filtering documents as part of the front end of our biomedical and clinical natural language processing system, BioMedICUS[3]. We hypothesized that DO would provide an adequate framework for document representation. Another hypothesis was that the intricacies of a vendor-supported EHR system with large numbers of document value sets would likely reveal certain incongruent or inconsistent aspects with the DO framework.

Background
Several researchers have explored the application of DO, mostly with document names in the inpatient setting. In a seminal study[4], Hyun *et al.* evaluated the adequacy of three versions of the DO to represent document names in the inpatient setting at Columbia University Medical Center. The authors found DO version 3 was superior to previous versions on both the level of specificity and completeness in document names as well as level of granularity of DO. Axes *SMD* and *TOS* in version 3 had value sets that needed expansion for better representation of document names.

To understand the DO in further depth and the effect of precoordination of its axes with LOINC, Dugas *et al.*[5] evaluated the coverage of LOINC codes on 86 document types from an inpatient hospital information system.

Similar to Hyun's work, the authors reported that more specific LOINC codes were necessary for better coverage. Subsequently, Chen *et al.*[6] explored the process of mapping document names from two large institutions to DO and identifying LOINC codes based on the mapping to further define the strengths and limitations of DO and LOINC codes for document representation. The results showed that a majority of document names can be assigned to a LOINC code with existing DO axis values but that there was often loss of information and granularity mismatch for one or more axis(es). In this study, mappings were performed for document names from the inpatient setting at one site and document attributes from a legacy system at the second site. Li *et al.*[7] investigated the coverage of DO on clinical document titles again in the inpatient setting from two campuses of the NewYork-Presbyterian Hospital and explored the process of using LOINC codes for exchange of documents across institutions. Similar coverage of DO and issues identified in the aforementioned studies were reported. The authors explored different combinations of DO axes for LOINC code mapping. As a small percent of local documents were reported to have an exact match to existing LOINC codes, the study showed that using LOINC codes with precoordination for document exchange might not be feasible or allow for maximal flexibility of document reuse.

In a study looking at nursing documents and their associated section headings, Hyun and Bakken[8] extracted section headings from nursing documents, identified DO components and mapped them to the LOINC semantic model. The study reported that 38% of the headings were successfully represented. The authors also found that in order to better represent nurse document components, values of the attributes of the LOINC semantic model needed to be extended. Finally, Shapiro *et al.*[9] analyzed the single DO axis of *SMD* on a set of document titles within the Medical Entities Dictionary at NewYork-Presbyterian Hospital. The study showed that 56% of document titles were classified as "not specified" on the DO *SMD* list. In the study, a new polyhierarchical SMD structure was created combining the values from the DO database with values from the American Board of Medical Specialties (ABMS). The resulting new structure significantly increased the coverage of *SMD* on document titles.

Methods

This study involved analysis, collection, text processing, and mapping of clinical documents stored in the clinical data repository from University of Minnesota-affiliated Fairview Health Services. The repository contains documents between 1993 and 2013 from a single electronic health record system (Epic™), as well as legacy documents from affiliate clinics for variable time periods. Fairview includes seven hospital sites and over 330 ambulatory clinic locations. Currently the repository hosts more than 66 million notes and the average daily document volume includes 133,000 document updates and 45,000 new note insertions.

Repository data items related to patient encounters, departments, providers and clinical documents were inspected to find items with values for each DO axis. Figure 1 shows data items such as "Encounter Type", "Position Type" collected from patient encounters. Each of these data items was compared to different axes of the DO and was analyzed to understand if each of the data items specified information in one or more DO axis(es). While the DO has at minimum population of the *Kind of Document* and at least one other DO axis, the distribution of axis population was analyzed, in order to understand the consequences of filtering by one or more of the DO axes based on analysis of the items and populating rates of each data item collected from the data repository. As illustrated in Figure 1, entries of department specialty, provider specialty and hospital service were mapped to the *SMD* axis.

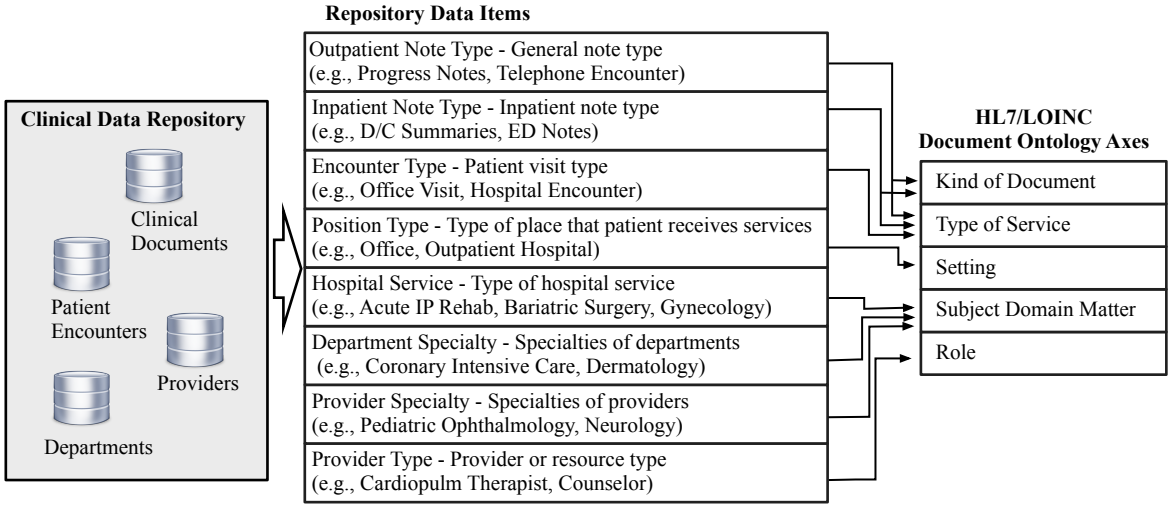


Figure 1. Mapping from Epic items to HL7/LOINC DO axis.

Table 1. Mapping examples of different data items.

Example	Document Ontology Axis(es) Mapping(s)					
Anesthesia Pre-op Evaluation (Inpatient Note Type)	KOD	7. Note	TOS	9.1.2. Preoperative Evaluation and Management	SMD	4.Anesthesiology
Consult/Results-Findings (Outpatient Note Type)	KOD	7. Note	TOS	3. Consultation		
Nursing Facility (Position Type)	Setting	6.b. Nursing Facility				
Palliative Care (Hospital Service)	SMD	30. Palliative Care				
Pediatric Neurology (Department Specialty)	SMD	19. Neurology				
Ent-Otolaryngology (Provider Specialty)	SMD	29.Otolaryngology				
Cardiopulm Therapist (Provider Type)	SMD	14.b. Cardiovascular Disease	Role	15. Therapist		
Prenatal Office Visit (Encounter Type)	TOS	7.b. Office				

Depending on the type of the visit (e.g., inpatient or outpatient) that a clinical note is associated with, different sets of data items can be used to extract information for DO axes. For example, for an inpatient encounter, the clinical document specialty (DO axis of *SMD*) usually can be extracted from the "Hospital Service", "Provider Specialty" or document titles. For an outpatient encounter, the SMD is often specified in the "Department Specialty", "Provider Specialty" or document titles. A list of data items, such as "Outpatient Note Type" and "Hospital Service", in the data repository were mapped to DO axes. Table 1 shows examples of mapping from data items to DO axes.

We adopted the same mapping and rating process described by Hyun *et al.* and Chen *et al*[4, 6]. Each data item was examined and values for applicable DO axes extracted from the entry. For example, from encounter type "Oncology Visit", DO *SMD* axis is extracted as "14.f. Hematology and Oncology". If an entry included related information to a DO axis, but no appropriate value was found in the existing value list, the entry was classified as "Not Covered" for that particular DO axis. If an entry contained no information for an axis of the DO, a value of "Not Specified" was assigned to the entry. A rating (*adequate*, *too broad*, *too specific*, *not covered or not specified*) was used to indicate the coverage of the particular DO axis values to each of the above mapping results. Inter-rater reliability was calculated using mappings of two reviewers on approximately 10% random subsets of data item entries.

Results

Table 2 contains a list of data items, the number of item entries in the repository, number of entries that specify values for each DO axis, and populating rate of each data item. Entries were found to sometimes specify information for more than one DO axis. For instance, encounter type "Anesthesia Consult" indicates both the *TOS* and *SMD*. Values for each DO axis can be obtained in several data items. For instance, some item entries, such as with an encounter type of "Case Management" and Department Specialty of "Dialysis", can all indicate the *TOS* axis. Information on *SMD* was contained to some extent in all data items.

Table 2. Distinct data item mappings to DO axis values.

Data Item	Entries	KOD	TOS	Setting	SMD	Role	Populating rate
Inpatient Note Type	90	83	79	7	3	3	100% (Inpatient)
Outpatient Note Type	65	49	46	5	2	2	96.5% (Outpatient)
Position Type	50	0	0	47	0	0	74.3/98.3% (Outpatient/Inpatient)
Hospital Service	98	0	6	14	87	5	91.2% (Inpatient)
Department Specialty	95	0	2	5	82	8	26% (Outpatient)
Provider Specialty	176	0	2	5	155	28	-
Provider Type	79	0	0	0	38	71	67%/76.5% (Outpatient/Inpatient)
Encounter Type	172	57	64	40	14	6	84.9%/100% (Outpatient/Inpatient)

The distribution of populating rates of each item hosted in the data repository (with the exception of those most related to provider specialty) is also summarized in Table 2. As shown in Table 2, some data items were well populated (>80%), including encounter type, hospital service and position type; whereas others, such as the department specialty and provider type, were fairly populated. For example, 47 out of 50 entries for "Position Type"

contain information about DO axis *Setting* and 30 of them can be mapped to an existing value of DO axis *Setting*. Table 3 shows proportions of finalized inpatient and outpatient notes populated with data item entries containing DO information and data item entries can be mapped to existing DO values. 95.3% of inpatient notes are populated with "Position Type" from the 47 entries and 91.4% of inpatient notes are populated with a "Position Type" from the 30 entries. The distribution of mapping ratings for data items and inter-rater reliability of mappings are summarized in Table 4.

Table 3. Inpatient and outpatient note population rates with data item entries containing DO information and with data item entries can be mapped to existing DO values.

	Inpatient Notes n=2,134,945 (16.57%)	**Outpatient Notes n=10,751,838 (83.43%)**
KOD	100.0% / 100.0% (Inpatient Note Type)	96.5% / 96.5% (Outpatient Note Type)
TOS	100.0% / 34.7% (Inpatient Note Type) 100.0% / 100.0% (Encounter Type)	96.5% / 96.5% (Outpatient Note Type) 65.9% / 65.9% (Encounter Type)
Setting	95.3% / 91.4% (Position Type)	71.8% / 71.3 % (Position Type)
SMD	86.9% / 85% (Hospital Service) -/- (Provider Specialty)	16.3% / 15.8% (Department Specialty) -/- (Provider Specialty)
Role	76.2% / 76.2% (Provider Type)	59.6% / 59.6% (Provider Type)

The majority of the data item entries contain DO axis information. Overall, existing values of DO axes are either adequate or too broad for data item entries that contain DO axis information. Existing *KOD* and *SMD* values can exactly specify most of the data item entries that contain *KOD* and *SMD* information. A number of new *Setting* types such as "Community Mental Health Center" and "Independent Laboratory" as well as new *Role* types such as "Athletic Trainer" and "Diabetes Educator" were discovered from the mappings. A large number of data item entries contain more specific *TOS* and *Setting* information than existing DO values for these two axes. For example, position type "End Stage Renal Disease Treatment Facility" is mapped to a less specific *Setting* value "7. Outpatient". Outpatient note type "Consult/Results – Findings" and "Consult/Results – Impression" are mapped to the same less specific *TOS* value "3. Consultation".

Table 4. Mapping ratings by axis and inter-rater reliability.

Mapping ratings (Inpatient setting=shading; Outpatient setting=no shading)					
	KOD	**TOS**	**Setting**	**SMD**	**Role**
Adequate	87.8%	20.6%	26.0%	58.0%	29.1%
	72.3%	13.1%	26.0%	58.6%	29.1%
Too Broad	3.3%	30.1%	32.0%	11.1%	24.0%
	1.5%	27.8%	32.0%	10.3%	24.0%
Too Specific	0	0.4%	2.0%	0.8%	0
	0	0	2.0%	0.7%	0
Not Covered	1.1%	3.4%	34.0%	5.4%	36.7%
	1.5%	5.5%	34.0%	5.8%	36.7%
Not Specified	7.8%	45.4%	6.0%	24.6%	10.1%
	24.6%	53.6%	6.0%	24.6%	10.1%
Inter-Rater Reliability on SMD axis					
Proportion Agreement	88.9%	92.3%	80.0%	84.6%	85.7%
	100%	91.3%	80.0%	88.5%	85.7%

*Mapped item rates of SMD is reported for inpatient documents mapping using *hospital service* data only and for outpatient using *department specialty* only.

Discussion

In this paper, we have described an analysis applying the HL7/LOINC DO for representing documents in a large research clinical data repository for a sizable integrated health system. We studied the structure of entities related to encounters, notes, providers and departments. A set of data items were collected and then mapped to the DO axes. Database populating rate of all data item entries that mapped to a particular DO axis were calculated and analyzed for both inpatient and outpatient settings. Mapping results showed that for both inpatient and outpatient documents, the majority of the related repository item entries can be mapped to a value in the defined list of the respective DO axis. Similar to previous studies on the adequacy of DO for clinical document names, we observed similar issues such as granularity issues and loss of information.

Further analysis on the data repository shows that most inpatient notes were populated with data item entries that are mapped to existing *KOD* and *Setting* types in HL7/LOINC DO. However, only 34.7% of inpatient document was populated with data item entries that can be mapped to existing *TOS* values. Upon inspection of documents in the data repository, we found that nearly half of the inpatient documents were populated with an "Inpatient Note Type" or "Miscellaneous", which cannot be mapped to an existing *TOS* value. The populating rate of hospital service that can be mapped to DO axis *SMD* was high (86.9%). Only 59.6% of the outpatient documents are populated with provider types that can be mapped to DO axis *Role* because of the low populating rate of the provider type at the institutional level, particularly for interdisciplinary staff. Also, only 16.3% of the outpatient documents are populated with mapped department specialties, which could be used for extracting information for DO axis *SMD*.

In addition to structured data items related to *SMD*, titles of clinical documents, such as "Sleep Medicine Chart Note", "FINAL PULMONARY CONSULTATION", and "AMB Nurse Triage Note", also contain *SMD* information. For instance, the document title "Sleep Medicine Chart Note" indicates an *SMD* value "19. Neurology" and title "FINAL PULMONARY CONSULTATION" indicates an *SMD* value "14.i. Pulmonary Disease". To better utilize information encoded in document titles, algorithms and tools need to be developed in future studies.

The University of Minnesota clinical data repository stores clinical documents from the current Epic system, as well as documents from legacy EHRs (i.e., AllScripts and Eclipsys) over a sustained period of time. The process of adding documents from a number of legacy systems has resulted in inconsistencies between data item entries. For instance, analysis of documents within the repository shows that a large number of legacy documents were populated with encounter type "Admission H&P", but associated with an outpatient note type "Progress Note".

The results presented show that the HL7/LOINC DO is able to represent a majority of clinical data repository data items. Similar issues such as granularity and loss of information were found in this study as reported in previous studies. Further effort is needed to develop tools to acquire with higher fidelity *SMD* and *TOS* for documents. We plan a more detailed analysis of *Role* and *Setting*, along with providing detailed mappings to the DO of these values.

Acknowledgements
The authors would like to thank Fairview Health Services and grant support from National Library of Medicine 1R01LM011364-01 (EC/GM), Agency for Healthcare Research and Quality 1R01HS022085-01 (GM), National Institute of General Medical Sciences 1R01GM102282-01A1 (SP), and the University of Minnesota Clinical and Translational Science Award 8UL1TR000114-02.

References

1. MacKenzie SL, Wyatt MC, Schuff R, Tenenbaum JD, Anderson N. Practices and perspectives on building integrated data repositories: results from a 2010 CTSA survey. Journal of the American Medical Informatics Association: JAMIA. [Research Support, N.I.H., Extramural]. 2012 Jun;19(e1):e119-24.
2. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. Journal of the American Medical Informatics Association: JAMIA. 2006 Jan-Feb;13(1):30-9.
3. BioMedICUS. Available from: http://code.google.com/p/biomedicus/.
4. Hyun S, Shapiro JS, Melton G, Schlegel C, Stetson PD, Johnson SB, et al. Iterative evaluation of the Health Level7--Logical Observation Identifiers Names and Codes Clinical Document Ontology for representing clinical document names: a case report. Journal of the American Medical Informatics Association: JAMIA. [Research Support, N.I.H., Extramural]. 2009 May-Jun;16(3):395-9.
5. Dugas M, Thun S, Frankewitsch T, Heitmann KU. LOINC codes for hospital information systems documents: a case study. Journal of the American Medical Informatics Association: JAMIA. 2009 May-Jun;16(3):400-3.
6. Chen ES, Melton GB, Engelstad ME, Sarkar IN. Standardizing Clinical Document Names Using the HL7/LOINC Document Ontology and LOINC Codes. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2010;2010:101-5.
7. Li L, Morrey CP, Baorto D. Cross-mapping clinical notes between hospitals: an application of the LOINC Document Ontology. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2011;2011:777-83.
8. Hyun S, Bakken S. Toward the creation of an ontology for nursing document sections: mapping section names to the LOINC semantic model. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. [Research Support, N.I.H., Extramural]. 2006:364-8.
9. Shapiro JS, Bakken S, Hyun S, Melton GB, Schlegel C, Johnson SB. Document ontology: supporting narrative documents in electronic health records. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. [Research Support, N.I.H., Extramural Validation Studies]. 2005:684-8.