Building a Richly Connected and Highly Analyzed Genotype/Phenotype Ecosystem in a World of Data Silos

Daniel T. Heinze, PhD[1], Scott Kahn, PhD[2], Paul A. McOwen, MS[1], Joel L. Vengco, MS[3], Elizabeth A. Worthey, PhD[4]

[1]Zato Healthcare, Easthampton, MA; [2]Illumina, San Diego, CA;

[3]Baystate Health, Springfield, MA; [4]Medical College of Wisconsin, Milwaukee, WI

Abstract

The ability to index, aggregate, search, navigate, analyze and share genomic and clinical data across departmental, institutional, geographic and political boundaries while maintaining security, privacy and data rights is critical to the success of translational medicine. We discuss advances in the technology of cooperative computing, information fusion and surface form ontologies with application to the translation of genomic research to clinical practice and, conversely, the application of phenotypic data to genomic research. Specifically, we describe a seminal collaboration of genomic R&D with clinical medicine as facilitated over a secure, clinically appropriate, ontology enabled, multi-centric platform for discovery across diverse genomic and clinical data sets that are stored and administered on diverse and disparate data centers and data types.

This environment motivates the investigation of a variety of genotype/phenotype issues. We discuss the migration in the clinical context from sparse phenotype information to fully extracted phenotype data from the full clinical record to ontologically structured phenotype data. In the genetic research context, we discuss the migration toward a structure of deeply analyzed and organized clusters of genotype/phenotype data.

Overview

The issues discussed by this panel are motivated by a newly introduced solution to the problem of sharing, analyzing and compositing genotype and phenotype data that is collected by hospitals, clinics, and researchers around the world. We bring together experts in the fields of clinical genomics, genetic research, and clinical information management as united by a medical ontology enabled cooperative computing environment that has proven itself in intelligence applications for national security. The Medical College of Wisconsin is a leader in the field of Genomic Medicine and has thus been exploring various clinically appropriate methodologies to solve the genotype and phenotype data sharing problem. Zato's software platform, represented by Paul McOwen, has been praised for successes from extensive operational use. Additional government sponsored and IBM sponsored performance testing at massive scales has demonstrated a unique capability for spanning massive data silos across departmental, institutional, governmental, and geographic boundaries as well as across the boundaries of network types, operating systems, database systems, file systems and file types (both structured and unstructured) without moving data from its original location and with full security and control by the data owners. Zato's clinical ontology, represented by moderator Daniel Heinze, provides the essential capability to combine both logical ontology and surface form ontology in a manner that enables high accuracy phenotype information extraction from indexed clinical records and methods for retrieval of relevant genotype data. Zato is working with: Baystate Health, represented by Joel Vengco, to index the complete Baystate Health clinical records systems; Illumina to index sequenced genome data; and the Medical College of Wisconsin to explore capabilities to enable researchers and clinicians around the globe to share sensitive patient information that will translate to improved clinical care based on the field of genomic medicine. In the environment that will be discussed, all participants locally host and control their own data while also, at each data owner's discretion, securely share, search, retrieve, extract from and process each other's data in a clinically appropriate manner to improve a patient outcome.

The underlying Zato platform is broadly deployed for intelligence and national security use and has proven unique, secure and scalable. It has been deployed at the Genomic Medicine program at the Medical College of Wisconsin and is in the process of being deployed at Baystate Health as well as in other clinical venues. Within the biomedical and clinical venues, the platform is seamlessly integrated with ontologies, information extractors and clinical applications developed by Zato.

Panelist Presentation Summary

Daniel Heinze (Moderator): Introduction**.**

Scott Kahn: Toward a Rich Genotype/Phenotype Ecosystem: This presentation will introduce the issues involved in sharing genomic data across the physical, spatial, geographic, institutional and political boundaries that create the data silo problem and hinder the development of rich, highly analyzed genotype information repositories. Further detail will be given to the matter of what constitutes a rich and highly analyzed genotype ecosystem, its benefits and steps toward creating such an ecosystem. Interaction will be solicited from the audience regarding experiences with consolidating and sharing genotype data and the challenges of consolidation and analysis across silos.

Paul McOwen: From Silos to Ecosystems: The problems created by data silos are elaborated and a unique and highly successful solution from the intelligence, defense, and national security communities is presented. The issues of data search and sharing while also maintaining local data control and storage, access authorization and revocation, and security will be detailed. Case studies will be presented demonstrating the scalability and effectiveness of the solution. Interaction will be solicited from the audience, particularly with regard to plans and methods for large-scale dissemination of the technology, tightly integrated with a comprehensive medical ontology, in the biomedical and clinical communities.

Joel Vengco: Toward a Rich Genotype/Phenotype Ecosystem – Part II: This presentation continues the theme related to the problems involved in consolidating and sharing siloed information, particularly as related to the clinical setting. Particular attention will be given to the need to evolve a rich and highly analyzed ecosystem in which the relations between individual and aggregate patient genetics, clinical observations, and treatment can be used to further the goals of research, development, and disease treatment and management. Interaction will be solicited from the audience particularly with regard to the suitability of the described ecosystem platform for insertion into the clinical information technology setting and discussion of its performance in the early-adopter setting.

Elizabeth Worthey: Genotype/Phenotype Ecosystem In Action: A use case will be presented describing an early-adopter collaboration and implementation across the physical, spatial, geographic and institutional boundaries of biomedical and clinical organizations. Particularly, an implementation by the Human and Molecular Genomics Center of the Medical College of Wisconsin with collaborative links to Baystate Health to demonstrate the feasibility and practical clinical implementation of shared Genotype/Phenotype Ecosystem will be demonstrated. Interaction will be solicited from the audience particularly with regard to discussion of the value proposition and feasibility of large-scale clinical dissemination of the model.

Dissemination of Pharmacogenomic Knowledge: Establishing a Pathway to Support Clinical Implementation

James M. Hoffman, PharmD, MS;[1] Michelle Whirl-Carrillo, PhD;[2] Josh F. Peterson, MD, MPH;[3] Robert R. Freimuth, PhD[4]

[1]St. Jude Children's Research Hospital, Memphis, TN, USA; [2]Stanford University, Palo Alto, CA, USA; [3]Vanderbilt University School of Medicine, Nashville, TN, USA; [4]Mayo Clinic, Rochester, MN, USA

Abstract

Pharmacogenomics is often an initial focus for the implementation of genomic medicine. To facilitate the translation of pharmacogenomic knowledge to clinical practice, authoritative resources are needed to form a knowledge base that combines genomic and medication information, which can be used to support gene-based prescribing through the electronic health record (EHR). As these resources are established, the knowledge must be represented in ways that will enable broad dissemination. This session will illustrate how three national initiatives are working together to establish a pathway to support the dissemination and clinical implementation of pharmacogenomics knowledge. Specifically, this panel will highlight the development of clinical guidelines by the Clinical Pharmacogenetics Implementation Consortium (CPIC), including its increased focus on informatics, and the dissemination of that knowledge through PharmGKB. The panel will also summarize lessons learned by the Pharmacogenetics Research Network (PGRN) Translational Pharmacogenetics Program (TPP), which identifies barriers for the implementation of pharmacogenomics, including integration with the EHR, and compares implementation approaches across diverse sites. Finally, because pharmacogenomic expertise may be concentrated in specific organizations and the technical architecture of clinical information systems varies widely, standards must be developed to share pharmacogenomic knowledge, especially genotype interpretations and prescribing recommendations.

Panel Overview

Pharmacogenomics is a frequent starting point for the implementation of genomic medicine. The number of clinically-actionable pharmacogenetic variants is steadily increasing, and a growing number of academic medical centers have invested substantial effort to implement pharmacogenomics knowledge. However, to broadly incorporate pharmacogenomics into clinical practice and electronic health records (EHRs) across all settings of care, pharmacogenomic knowledge must first be centralized by authoritative resources. As these resources are established, the knowledge must be represented in ways that enable broad dissemination.

This session will illustrate how three national initiatives establish a pathway to support the broad dissemination and clinical implementation of pharmacogenomic knowledge. First, resources and tools from PharmGKB, a longstanding and well-recognized knowledge base for pharmacogenomics, will be summarized. Next, the activities of the Clinical Pharmacogenetics Implementation Consortium (CPIC), which is a shared project of Pharmacogenetics Research Network (PGRN) and PharmGKB, will be reviewed. As of October 2013, CPIC has produced 10 guidelines that are designed to help clinicians understand how available genetic test results should be used to optimize drug therapy. More recently CPIC has increased its focus on informatics by forming a working group (CPIC Informatics) that supports the adoption of the CPIC guidelines by identifying, and resolving where

possible, potential technical barriers to the implementation of the guidelines within a clinical electronic environment. CPIC Informatics has started by developing comprehensive translation tables that illustrate how genotype test results can be used to infer molecular phenotype, which is used as the basis for clinical recommendations that can be implemented as clinical decision support (CDS) within an EHR. Third, the panel will summarize lessons learned by PGRN's Translational Pharmacogenetics Program (TPP), a group of eight sites that are independently implementing pharmacogenomic CDS rules, specifically identifying barriers to the implementation of pharmacogenomics, including integration with local EHRs. In addition, the TPP allows comparison of implementation approaches across diverse sites through a set of structured documents that provide unique insights on the resources that are required to implement pharmacogenetics in the clinical environment.

Finally, because not all settings of care may have adequate expertise in pharmacogenomics and because the technical architecture of clinical information systems varies widely, standards must be developed to share pharmacogenomic knowledge, including genotype interpretations and prescribing recommendations. Various methods to share pharmacogenomic knowledge will be highlighted throughout the panel. Recent and ongoing national initiatives will be reviewed.

The objectives of this panel will be to:

- Describe two authoritative resources for pharmacogenomic knowledge (PharmGKB and CPIC)
- Illustrate how these resources can be used to aid the clinical implementation of pharmacogenomics
- Identify lessons learned when the implementation of pharmacogenomics is coordinated across diverse settings, as exemplified by the experiences of the PGRN TPP
- Review ongoing and future directions for sharing pharmacogenomic knowledge
- Discuss how the pathway defined by these initiatives supports the clinical implementation of pharmacogenomics and may establish best practices for the implementation of genomic medicine beyond pharmacogenomics.

This panel will follow the recommended format of 4 brief presentations with ample time for questions and discussion with the audience. Dr. Hoffman will coordinate planning across all presentations to present a cohesive panel, and he will act as the panel's moderator. The presenters already work together regularly, which will facilitate the integration of presentations expected for AMIA TBI panels. Also, all presenters have experience in multiple topics on the panel, which will facilitate discussion among the panelists and with the audience. For example, while Dr. Freimuth's presentation will focus on TPP, all panelists participate in TPP so multiple perspectives will be presented in the discussion. Finally, this panel was planned in coordination with AMIA genomics working group (Dr. Freimuth, Chair), and it is aligned with this Working Group's interest in implementing genomics into the EHR.

Michelle Whirl-Carrillo – Speaker; Stanford University

Michelle Whirl-Carrillo, PhD, is the Assistant Director of PharmGKB in the Department of Genetics, Stanford University. She has been with the PharmGKB project for over ten years, with a brief absence to start the pharmacogenomics effort at a direct-to-consumer genotyping company. Dr. Whirl-Carrillo has an S.B. in Biology from MIT and a Ph.D. in Biophysics from Stanford University. She has extensive experience annotating, evaluating and interpreting a broad range of pharmacogenomic associations from preliminary findings to clinically actionable results. Dr. Whirl-Carrillo is an active member of the Clinical Pharmacogenetics Implementation Consortium

(CPIC), a co-leader of the CPIC Informatics working group and an author on several of the CPIC guidelines. She is also involved the PGRN Translational Pharmacogenomic Project (TPP) working group. She will discuss PharmGKB's role in the annotation and dissemination of pharmacogenomic information.

James Hoffman – Speaker and Moderator; St. Jude Children's Research Hospital
James Hoffman, PharmD, MS is an Associate Member in Pharmaceutical Sciences and St. Jude's Medication Outcomes and Safety Officer. He has been a member of the Clinical Pharmacogenetics Implementation Consortium (CPIC) since its inception. Along with others on the panel, he is a co-leader of the new CPIC Informatics working group. Dr. Hoffman is also an investigator on St. Jude's protocol PG4KDS protocol to implement preemptive pharmacogenomics (www.stjude.org/pg4kds), which has relied on the resources provided by PharmGKB and CPIC. In his presentation, he will review the CPIC guideline development process and summarize the priorities and initial work for CPIC informatics, which was formed in 2013. Dr. Hoffman will also act as the moderator for this panel.

Robert Freimuth – Speaker; Mayo Clinic
Robert Freimuth, PhD, is an Assistant Professor in the Department of Health Sciences Research, Mayo Clinic. He is a leader in the Pharmacogenomics Ontology (PHONT) PGRN network resource, which aims to help standardize pharmacogenomic data representations. He leads the Data Standardization Working Group within the PGRN Translational Pharmacogenomic Project (TPP) and he is a co-leader of the Clinical Pharmacogenetics Implementation Consortium (CPIC) Informatics Working Group. He is currently working with the Mayo Clinic Center for Individualized Medicine and eMERGE PGx to implement pharmacogenomic CDS rules within the Mayo EHR system. Dr. Freimuth is developing a formal knowledge representation based on national standards to enable the expression and sharing of pharmacogenomics guidelines. He will discuss challenges and barriers to developing and implementing CDS for pharmacogenomics, including the experiences of the TPP and issues related to locus-specific allele nomenclature. Dr. Freimuth is the current Chair of the AMIA Genomics Working Group.

Josh Peterson – Speaker; Vanderbilt University Medical Center
Josh Peterson, MD, MPH is an Assistant Professor of Biomedical Informatics and Medicine at Vanderbilt University Medical Center. He has developed, implemented and evaluated CDS projects to personalize therapy for the elderly, for patients with compromised kidney function, and most recently, for patients with variant drug metabolism genes. He is a leader in the effort to implement and evaluate Vanderbilt's PREDICT program, which has incorporated routine pharmacogenomics testing at the point of care for greater than 14,000 patients since late 2010. He will describe the key lessons of the PREDICT implementation, and describe Vanderbilt's effort to disseminate methods and tools relate to PREDICT by participating with the Translational Pharmacogenomic Project (TPP) and eMERGE PGx, the pharmacogenomic implementation study of the eMERGE program. Additionally, Dr. Peterson will describe the prospects for translating knowledge that is formalized by CPIC and PharmGKB into remote EMRs through the use of web services.

Affirmation and Potential Additional Support

All proposed panel members have been personally contacted by James Hoffman, and they have agreed to participate in this panel. Russ Altman, who is the current president of the American Society for Clinical Pharmacology and Therapeutics (ASCPT) is interested in promoting the intersection of clinical pharmacology and medical informatics. He has reviewed this proposal, and if the panel is approved, Dr. Altman would like to identify ways for ASCPT to advertise support this panel (and perhaps AMIA TBI in general) to ASCPT members.

Strategies for Sustainable Open Source Projects for Clinical and Translational Research: Lessons from the Trenches

Elizabeth K. Nelson, PhD[1] (Moderator), Leon Rozenblit, PhD, JD[2], Michael Mendis, BS[3], Ben Bauman, BA[4], Mark Igra, BS[1]
[1]LabKey Software, Seattle, WA; [2]Prometheus Research, New Haven, CT; [3]Harvard Partners, Boston, MA; [4]OpenClinica, Waltham, MA;

Abstract

Theoretically, taking an open source approach can broaden the public benefits of grant-funded software projects; increase the leverage of informatics investments; draw upon a wider pool of contributors and expertise; and improve transparency, reproducibility, and extensibility. However, as Dr. Isaac Kohane has warned, open source software is "…free like a pony. You still have to feed it and clean up after it" (TEDMed, 2013). Furthermore, simply making software open source does not ensure that it will become immediately useful to others. This panel will cover practical strategies for generalizing, sustaining, and evolving open source software developed for clinical and translational research. Panel members will address sustainable business models, feasibility of grant support, implications of different open source licenses, modes of dissemination (including community norms for attracting open-source evangelists), community-building approaches, practical trade-offs, and unexpected challenges. Panelists represent open source platforms for clinical and translational research that have proven useful across multiple organizations and shown sustainability over time. Platforms include LabKey Server (http://labkey.org), RexDB (http://rexdb.org/), i2b2 (https://i2b2.org/), and OpenClinica (https://openclinica.com/).

Background

"i2b2 is free and open source. Free like a pony. You still have to feed it and clean up after it.

- Isaac Kohane, TEDMed Conference, Washington DC, April 19, 2013

"Open source has the half life of a graduate student."

- Don Listwin, Xconomy Forum, Seattle, Washington, May 12, 2010

Despite the idealistic promise of open source software, actually delivering broad, lasting public benefit requires solid business strategy, long-term vision for software architecture, evangelism verging on mania, and gallons of elbow grease. Otherwise, as Don Listwin quips, an open source tool will have the "half life of a graduate student."

The increasing scope of clinical and translational research keeps raising the bar for informatics tools, both in the features delivered and the confidence investigators need in software longevity before considering adoption. For example, the longer the time frame researchers expect a study to continue, the longer software must be sustained, supported and extended to meet evolving needs.

For open source tools for clinical and translational research to become more widely accepted and adopted, the informatics community needs to better understand how to build, sustain and share open source projects in a way that supports field-specific needs and expectations.

To fulfill the promise of multiplying returns from informatics funding, open source tools need to be designed and developed in a way that meets the needs of a broader community, not just initial inventors. Furthermore, reusing proven open source, not reinventing the wheel, needs to become a norm for the field.

For biomedical open source projects to remain in good health (and cheerfully supported) after initial grants expire, the research community needs to explore and share business/funding models that have proven practical.

Panelists will address the value of open source approaches in clinical and translational research, the potholes that impede success, and strategies that have worked for their teams in establishing sustainable, widely used open source platforms.

Panelists and Presentations

Panelists represent open source platforms that have proven sustainable over time and useful across multiple clinical and translational research organizations. As of 2013, the software systems represent 29 collective years of experience as public open source projects. The platforms and their dates of first release are: LabKey Server (2005), RexDB (2005), i2b2 (2007), and OpenClinica (2005).

Each panelist will provide an overview of the open source project he/she represents, describe the project's "special sauce," and explore lessons learned. For ease of comparison, panelists will also address a predefined list of core points. The different platforms' strategies will be summarized incrementally on slides whose content are built across the presentations. Issues and metrics covered for each platform will include: (1) target users, (2) core value for users, (3) business/funding model, (4) challenges of funding model, (5) open source license, (6) number of active installations, (7) team size, (8) origins (academic or otherwise), (9) dissemination approach, and (10) community-building strategies. This will allow clear, consistent comparison of the platforms and strategies used to grow, maintain and evolve them, plus the tradeoffs made by each platform according to its focus.

Platform #1: LabKey Server

LabKey Server is an open source platform for large-scale, translational research.[1] The system helps teams of researchers collaborate smoothly and make sense of the flood of complex data produced by modern biomedical research, from novel assay and 'omics results to clinical reports to specimen information. LabKey Server supports web-based integration, analysis and secure sharing of diverse data types within distributed research teams. Upon research publication, the system can serve as a portal for public, interactive exploration of published analyses and de-identified data, opening doors to validation and extension of results.

Installations of the LabKey Server platform serve leading scientific organizations all over the world, including the Immune Tolerance Network (ITN TrialShare: http://itntrialshare.org), the Statistical Center for HIV & AIDS Research at the Fred Hutchinson Cancer Research Center (FHCRC) (Atlas: http://atlas.scharp.org) and NWBioTrust (http://www.nwbiotrust.org). There are currently over 100 active installations of the platform.

LabKey Server originated within the FHCRC, with its first public release in 2005. In the same year, the project's developers founded LabKey Software to support, extend, and sustain the platform beyond what was possible within a research institute. Today, research organizations use LabKey Server's rich API to develop new features independently or in partnership with LabKey Software. Code contributed by users (such as an electronic health record system for primate centers) undergoes code review by the LabKey Software team before addition to the source depot.

Groups that purchase LabKey Software's services apply a portion of this funding to maintenance of core platform infrastructure, including documentation. Successful dissemination and community-building strategies have included publishing papers, hosting a yearly user conference, hosting/supporting community message boards, and in-person outreach.

Challenges include funding core platform maintenance and installers; funding innovation that goes beyond current customer projects; growing the team quickly enough to meet user needs; simplifying the new user experience for a tool that serves a broad variety of researchers; and managing expectations for the cost of support for free software.

LabKey Server source code, compiled binaries, documentation, and tutorials are professionally maintained and freely available under the Apache 2.0 license at http://www.labkey.org.

Platform #2: RexDB

The Research Exchange Database (RexDB®) is an extensible, web-native software platform that helps researchers securely collect, integrate, manage and share data. RexDB embraces the flexibility required to support the needs of dynamic scientific collaborations. RexDB systems can be regularly updated to meet the changing needs of ongoing research projects, allowing many modifications to be performed by nontechnical software users. RexDB is built on top of sustainable open-source components, ensuring that research data will never be marooned in a proprietary format or a legacy system. A unique, layered architecture allows novice and advanced users, data managers, statisticians, and local IT staff to effectively interact with the system, while a highly granular privileging model ensures easy compliance with local and federal regulations.

RexDB is presently the data collection, data warehousing, and data sharing software platform for numerous longitudinal multidisciplinary research studies, including the Simons Foundation's Simons Simplex Collection and Variation in Individuals Project, the Yale Child Study Center's Yale Autism Research Database, and the centralized data solution at Emory University's Marcus Autism Research Institute. Dr. Rozenblit will address the following points.

- Users: Research organizations that generate data with high potential reuse value
- Sales pitch: Reduce the cost of data integration and repurposing across all your studies, data types and sites; transform your data into a value-generating asset.

- License strategy: Affero General Public License version 3 (AGPLv3), contributor covenants; special "permissive" licensing on high-value components
- Licensing challenges: managing tension between a for-profit's desire to limit competition and commitment to open source
- Funding model: NSF and NIH grants; professional open source, coupled with data management services, with service delivery methods accelerated by OS technology
- Funding challenges: Professional services model limits profit margins, which limits funding for product enhancement
- Open-source community building: Conferences (technical/scientific), Meetups
- Special sauce: Promoting and releasing high-value individual components as developer tools (e.g., HTSQL, cogs)
- Lessons learned: Tightly scoped developer tools are easier to promote than integrated systems.

RexDB is available at: http://www.rexdb.org/.

Platform #3: i2b2

i2b2 (Informatics for Integrating Biology and the Bedside) is a scalable informatics framework that enables clinical researchers to use existing clinical data for discovery research.[2] i2b2 is funded as a cooperative agreement with the National Institutes of Health (NIH).

Both open source communities and proprietary companies have benefited from the ability to add and extend the functionality of the i2b2 software by way of modular components called "cells." An example of an i2b2 cell would be a de-identification cell that strips names and dates from reports. Cells are exposed through web services, so proximity is not assumed. Remote cells can be hosted, so code does not have to be shared. For the open source community, i2b2 provides an infrastructure that consists of a wiki site, source code repository, and bug tracker. Two models are available for the development of other cells: a "Sponsored Projects" model, which provides the complete open source stack, and "Related Projects" model that includes a wiki site. As with all open source projects, i2b2 continues to push to engage the community and incorporate input and enhancements into the i2b2 platform. These efforts aim to broaden the community and improve the overall software. Within the last year, i2b2 has worked on running on a complete open source stack, with the last components being the use of an open source database.

i2b2 source code, virtual machines images, documentation and tutorials are available at http://www.i2b2.org under the i2b2 license.

Platform #4: OpenClinica

OpenClinica is an open source clinical trials data management platform. With thousands of implementations and a community comprising over 18,000 people, the software is used in over 100 countries across a diverse spectrum of both academic and industry research. OpenClinica helps increase the flow of clinical trial data, provides a more adaptable research IT infrastructure, and reduces the barriers to obtaining enterprise-quality electronic data capture and clinical data management systems.

Mr. Baumann will describe elements crucial to OpenClinica's success since the project's inception, recount certain things that have not worked as planned, and provide an overview of the OpenClinica community and commercial open source model that has allows it evolve into a sustainable initiative.

OpenClinica is distributed under the GNU Lesser General Public License (GNU LGPL) at https://community.openclinica.com.

Plan for Audience/Panel Interaction

The moderator will prepare questions for the panelists, but the focus of the interactive portion of this session will be on audience questions. In particular, this panel aims to help nascent open source projects strategize their own modes of sustainability and dissemination. The Q&A session will allow the audience to dig further into the varied strategies used by the open source projects represented on the panel.

Acknowledgements

We gratefully acknowledge grants from the NIH, including support for LabKey Server (UM1AI068618 and U01AI068635) and i2b2 (U54LM00874).

All coauthors have agreed to take part in the panel.

References

1. Nelson EK, Piehler B, Eckels J, et al. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. BMC Bioinformatics. 2011;12:71.
2. Murphy SN, Churchill SE, Bry L, et al. Instrumenting the health care enterprise for discovery research in the genomic era. Genome Res. 2009;360(13)1278-81.

EHR-based phenome wide association study in pancreatic cancer

Tomasz Adamusiak MD PhD[1*], Mary Shimoyama PhD[1,2]
[1]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI
[2]Department of Surgery, Medical College of Wisconsin, Milwaukee, WI
[*]tomasz@mcw.edu

Abstract

BACKGROUND. *Pancreatic cancer is one of the most common causes of cancer-related deaths in the United States, it is difficult to detect early and typically has a very poor prognosis. We present a novel method of large-scale clinical hypothesis generation based on phenome wide association study performed using Electronic Health Records (EHR) in a pancreatic cancer cohort.* ***METHODS.*** *The study population consisted of 1,154 patients diagnosed with malignant neoplasm of pancreas seen at The Froedtert & The Medical College of Wisconsin academic medical center between the years 2004 and 2013. We evaluated death of a patient as the primary clinical outcome and tested its association with the phenome, which consisted of over 2.5 million structured clinical observations extracted out of the EHR including labs, medications, phenotypes, diseases and procedures. The individual observations were encoded in the EHR using 6,617 unique ICD-9, CPT-4, LOINC, and RxNorm codes. We remapped this initial code set into UMLS concepts and then hierarchically expanded to support generalization into the final set of 10,164 clinical concepts, which formed the final phenome. We then tested all possible pairwise associations between any of the original 10,164 concepts and death as the primary outcome.* ***RESULTS.*** *After correcting for multiple testing and folding back (generalizing) child concepts were appropriate, we found 231 concepts to be significantly associated with death in the study population.* ***CONCLUSIONS.*** *With the abundance of structured EHR data, phenome wide association studies combined with knowledge engineering can be a viable method of rapid hypothesis generation.*

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act introduced the concept of Meaningful Use of information technology in health care. As part of this process, the legislation mandated the use of standard terminologies for electronic exchange of health information. Patient clinical records represent a largely untapped treasure trove of research information, which only recently has become more accessible thanks to the increasing adoption of Electronic Health Records and healthcare data standards. The need to integrate and exchange clinical data has long been recognized[1], but it was the HITECH Act that provided the final piece of the puzzle in terms of financial incentives.

A number of terminology standards are currently in use. **LOINC** (Logical Observation Identifiers Names and Codes) is a universal standard for identifying laboratory observations[2]. **RxNorm** is a standardized nomenclature for generic and branded drugs, as well as drug delivery devices. RxNorm provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software[3]. The Healthcare Common Procedure Coding System (HCPCS) maintained by the Centers for Medicare & Medicaid Services (CMS) is a standardized coding system for describing items and services provided in the delivery of healthcare[4]. It incorporates Current Procedural Terminology (**CPT**), a coding system maintained by the American Medical Association (AMA) to identify medical services and procedures furnished by physicians and other health care professionals[5]. International Classification of Diseases, Clinical Modification (**ICD-9-CM**) is an adaption created by the U.S. National Center for Health Statistics (NCHS) and used in assigning diagnostic and procedure codes associated with inpatient, outpatient, and physician office utilization in the United States[6]. All these terminologies are integrated within the **UMLS** (Unified Medical Language System) maintained by the National Library of Medicine (NLM)[7].

Current state of the art in extracting actionable information from EHR relies on large scale text-mining and NLP of clinical notes[8,9] or either focuses on a specific terminology within the EHR, e.g., ICD-9-CM[10,11] or looks into a handcrafted, small subset of EHR variables[12]. Our approach is novel in the sense that we analyzed the complete corpus of structured data within the EHR across all available terminology standards, as well as used an existing knowledge base (UMLS) to expand and generalize the findings.

Methods

Extract, Load and Transform (ELT)

A *Limited Data Set*, as defined under the Health Insurance Portability and Accountability Act (HIPAA), was obtained from the Medical College of Wisconsin Clinical Research Data Warehouse for this analysis. The data extract was in the form of standard Epic Clarity tables for a subset of patients that had an encounter or a problem list code in the *Malignant neoplasm of pancreas* (ICD9:157) code subset:

157 Malignant neoplasm of pancreas

- **157.0** Malignant neoplasm of head of pancreas
- **157.1** Malignant neoplasm of body of pancreas
- **157.2** Malignant neoplasm of tail of pancreas
- **157.3** Malignant neoplasm of pancreatic duct
- **157.4** Malignant neoplasm of islets of langerhans
- **157.8** Malignant neoplasm of other specified sites of pancreas
- **157.9** Malignant neoplasm of pancreas, part unspecified

Data was loaded into our in-house clinical analytics portal (ClinMiner), which was used to dynamically translate between any of the underlying clinical terminologies, and provided a consolidated view of the underlying patient data in a single UMLS perspective [13]. Drug information in EHR was encoded using MediSpan terminology, one of the RxNorm sources, which facilitated its automatic translation into UMLS. Labs were encoded as orders using CPT-4 codes or using a fixed category from the *CLARITY_COMPONENT* lookup table. We have manually mapped 130 tests from *CLARITY_COMPONENT* to LOINC, which provided coverage for over 97% of all lab observations (1 493 101 observations in total). Remaining ~3% lab observations were left unmapped and excluded from further analysis.

The source annotation space covered 6 617 unique ICD-9, CPT-4, LOINC, and RxNorm codes. This code set was then remapped into UMLS to facilitate further analysis, which resulted in 6 741 distinct UMLS CUIs (Concept Unique Identifiers). This code set was then expanded across a limited set of *is_a* and selected other relationships (e.g., *has_ingredient* for RxNorm drugs) as an extension of the method previously proposed in a method similar to that of parent child analysis described by Grossmann et al. [14].

but not beyond the original set of UMLS Metathesaurus semantic types of the expanded concepts to exclude functional concepts from the analysis and to keep the general meaning of the originating concept in the expansion. Additionally, the UMLS traversal was limited to either the UMLS Metathesaurus itself, or any of the following terminologies specific to Meaningful Use: RxNorm, NDF-RT, LOINC, SNOMED CT, HCPCS, and ICD-9-CM. This resulted in 18 038 concepts. Finally, we discarded 7 874 concepts that did not increase information content (i.e., were redundant in terms of partitioning of the underlying data) to reach the final 'phenome' of 10 162 concepts.

Statistical analysis

A chi-squared test was used to (χ^2) to test the significance of the associations. To correct for multiple testing we used a Bonferroni correction and tested at a level of $p < 4.9 \times 10^{-6} (0.05/10162)$. Odds Ratio (OR) and Relative Risk (RR) were used to assess the effect size of associations found to be significant.

Results

713 concepts were found to be significantly associated with death in the study population. Where both parent and its child concepts were found to be significant, child concepts were removed to further generalize the results and final

result set was thus reduced to 231 terms. A breakdown of all concepts by category and number of observations is shown in Figure!1.

Most of the terms were positively correlated with death and only the following 9 concepts were found to be associated with lower relative risk of death in the study population:

- Immunoassay for tumor antigen, quantitative; CA 125
- Vitamin D; 25 hydroxy, includes fraction(s), if performed
- Prealbumin measurement
- Racial group
- Benzoic acid or derivative
- Iodine AND/OR iodine compound
- Ionic iodinated contrast media
- Triiodobenzoic Acids
- sevoflurane Inhalant Solution

For practical reasons, only the top ten (five from each side) significant associations are shown in Table 1. The complete result set encompassing all 231 significant associations is available as supplementary materials at http://dx.doi.org/10.6084/m9.figshare.816958.

Discussion

As with any retrospective observation the primary limitation is a lack of a prospective control group, which means the results can be biased due to an imbalanced design. It is also worth noting, that correlation does not imply causation. For example, while cytopathology was found to be associated with an increased risk of death, it is more likely due to selection bias. Patients with more advanced disease more frequently underwent the procedure as part of their diagnostic process. There are also limitation due to data incompleteness. For example, here we looked at known deaths from the EHR only and did not include data from outside sources such the National Death Index. On the other hand, retrospective designs have the advantage of observing real clinical practice.

We have observed that the use of contrast media and medical gases used to induce anesthesia lowered the risk of death in the study population. This confirms an already known association between hospital resource utilization and patient mortality [15,16].

Cimetidine, an H2 receptor antagonist, has a known off-label use as an anticancer drug [17–19]. Paradoxically, its use was associated with an increased risk of death in our study population. However, this subpopulation was also older than the rest of the cohort, and likely increased mortality was due to a more advanced disease process. Without access to clinical notes, we can only speculate that perhaps this was a part of an experimental treatment.

We see the potential to use this approach to automatically generate groupers or value sets of closely related concepts. This could be used either in the EHR to alert the physician to other possibly relevant features of patient presentation as well as on the research side to make more informative patient cohort selections.

A major critique would be that they we only looked at association of the concept and not the value. Presence of an observation on a patient-level also discards the temporal and frequency information. On the other hand, this would also increase dimensionality of the analysis (cf. *curse of dimensionality*) and would require not only a more sophisticated statistical approach but could also suffer from lower statistical power. These are some of the challenges that we hope to address in future work.

Label	*CUI*	*Semantic Type*	Exposed Deceased	Exposed Alive	Not Exposed Deceased	Not Exposed Alive	p	OR	RR
			Increased Risk (RR > 1)						
Cytopathology, fluids, washings or brushings, except cervical or vaginal; smears with interpretation	*C0374051*	*Laboratory Procedure*	28	8	808	310	8.31×10^{-11}	9.12	**2.80**
Cimetidine	*C0008783*	*Pharmacologic Substance*	17	6	810	321	2.03×10^{-6}	7.14	**2.60**
Hyposmolality and/or hyponatremia	*C0020645*	*Finding*	22	10	806	316	6.54×10^{-7}	5.61	**2.44**
Osmolality; blood	*C0373690*	*Laboratory Procedure*	43	25	791	295	2.29×10^{-10}	4.61	**2.32**
Haptoglobin; quantitative	*C0373631*	*Laboratory Procedure*	25	14	802	313	1.17×10^{-6}	4.57	**2.28**
			Decreased Risk (RR < 1)						
sevoflurane Inhalant Solution	*C1253873*	*Clinical Drug*	7	85	731	331	1.90×10^{-6}	0.18	**0.24**
Triiodobenzoic Acids	*C0041013*	*Organic Chemical*	56	332	484	282	3.00×10^{-15}	0.28	**0.39**
Ionic iodinated contrast media	*C0361904*	*Indicator, Reagent, or Diagnostic Aid*	57	332	484	281	6.66×10^{-15}	0.29	**0.39**
Iodine AND/OR iodine compound	*C0303013*	*Inorganic Chemical Pharmacologic Substance*	60	338	478	278	1.38×10^{-14}	0.30	**0.40**
Benzoic acid or derivative	*C0578497*	*Organic Chemical*	58	328	488	280	4.41×10^{-14}	0.30	**0.41**

Table 1: Top ten events by effect size significantly associated with death in the study cohort. Abbreviations: CUI – Concept Unique Identifier; OR – Odds Ratio; RR – Relative Risk.

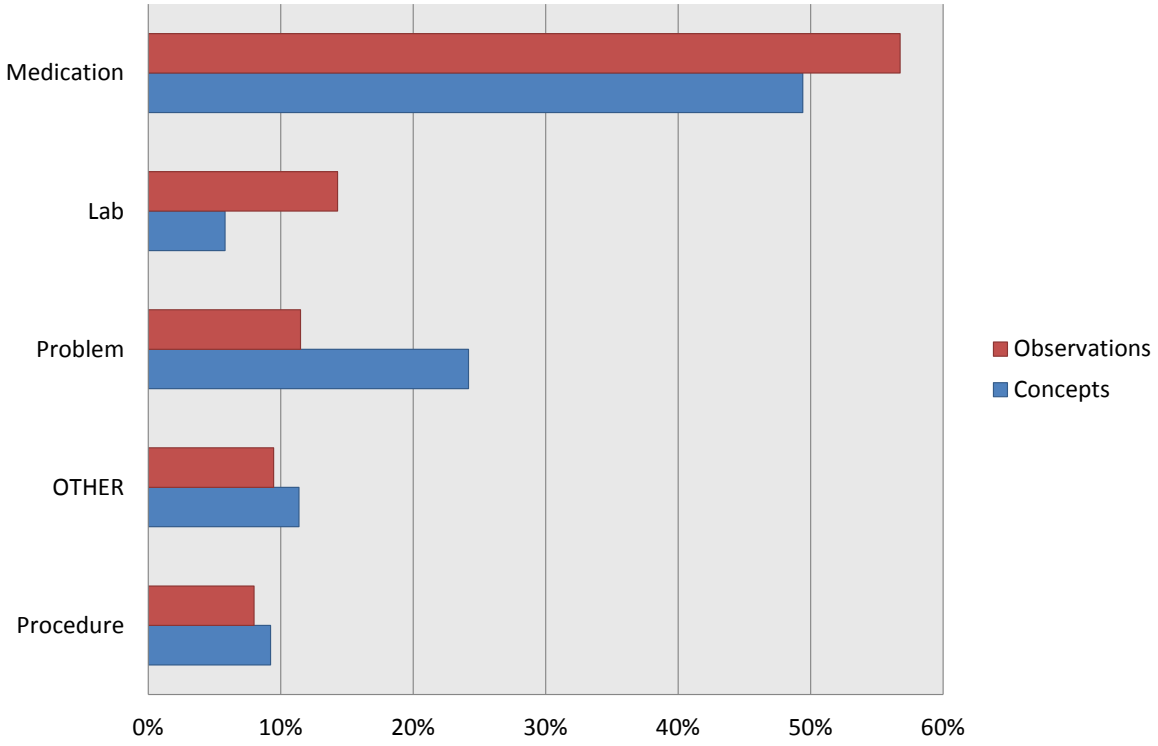


Figure 1: Breakdown of all 10 164 concepts by semantic type category and proportion of observations annotated with a particular concept. Categories are defined as follows. **Lab** is any UMLS concept in the semantic type tree: *A2.3.1. Clinical Attribute*, *A2.2.1 Laboratory or Test Result*, or *B1.3.1.1 Laboratory Procedure*. **Procedure** is a UMLS concept that is in the *B1.3.1 Health Activity* branch, but is not a *1.3.1.1 Laboratory Procedure*. **Problem** is a concept with a semantic type under *B2.2.1.2 Pathologic function* or *A2.2.2 Sign or symptom*. **Medication** groups all concepts classified by the UMLS Semantic Network either under semantic type *A1.4 Substance* or under *A1.3.3 Clinical Drug*. Finally, **OTHER** groups all other semantic types.

Conclusions

Information contained in EHR combined with knowledge engineering could be used a a viable method of rapid hypothesis generation, but requires comprehensive validation.

Acknowledgments

This project was funded in part by the Advancing a Healthier Wisconsin endowment at the Medical College of Wisconsin and the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through grant UL1 RR031973. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

We thank Stacy Zacher, Glenn Bushee, and Bradley Taylor for their help.

References

[1] J.J. Cimino and E.H. Shortliffe, editors. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine (Health Informatics)*. Springer-Verlag New York, Inc., Secaucus, NJ, 2006.

[2] C. J. McDonald. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clinical Chemistry*, 49(4):624–633, April 2003. ISSN 0009-9147. doi: 10.1373/49.4.624.

[3] Fola Parrish, Nhan Do, Omar Bouhaddou, and Pradnya Warnekar. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, page 1057, January 2006. ISSN 1942-597X.

[4] R Finnegan. HCPCS–outpatient procedures. *Journal (American Medical Record Association)*, 58(10):20–2, October 1987. ISSN 0273-9976.

[5] Current procedural terminology (CPT). *JAMA : the journal of the American Medical Association*, 212(5):873–4, May 1970. ISSN 0098-7484.

[6] R Finnegan. ICD-9-CM. *Journal (American Medical Record Association)*, 57(7):34–5, July 1986. ISSN 0273-9976.

[7] D A Lindberg, B L Humphreys, and A T McCray. The Unified Medical Language System. *Methods of information in medicine*, 32(4):281–91, August 1993. ISSN 0026-1270.

[8] Nicholas J Leeper, Anna Bauer-Mehren, Srinivasan V Iyer, Paea Lependu, Cliff Olson, and Nigam H Shah. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PloS one*, 8(5): e63499, January 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0063499.

[9] Svetlana Lyalina, Bethany Percha, Paea Lependu, Srinivasan V Iyer, Russ B Altman, and Nigam H Shah. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *Journal of the American Medical Informatics Association : JAMIA*, August 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2013-001933.

[10] Joshua C Denny, Marylyn D Ritchie, Melissa A Basford, Jill M Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R Masys, Dan M Roden, and Dana C Crawford. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics (Oxford, England)*, 26(9):1205–10, May 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq126.

[11] Jeremy L Warner, Amin Zollanvari, Quan Ding, Peijin Zhang, Graham M Snyder, and Gil Alterovitz. Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. *Journal of the American Medical Informatics Association : JAMIA*, August 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2013-001861.

[12] George Hripcsak and David J Albers. Correlating electronic health record concepts with healthcare process events. *Journal of the American Medical Informatics Association : JAMIA*, August 2013. ISSN 1527-974X. doi: 10.1136/amiajnl-2013-001922.

[13] Tomasz Adamusiak, Shimoyama Naoki, Tutaj Marek, and Shimoyama Mary. Next Generation Ontology Browser. In *Proceedings International Conference on Biomedical Ontology 2013*, pages 131–132, 2013.

[14] Steffen Grossmann, Sebastian Bauer, Peter N Robinson, and Martin Vingron. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics (Oxford, England)*, 23(22): 3024–31, November 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm440.

[15] Corrado Cecchetti, Riccardo Lubrano, Sebastian Cristaldi, Francesca Stoppa, Maria Antonietta Barbieri, Marco Elli, Raffaele Masciangelo, Daniela Perrotta, Elisabetta Travasso, Claudia Raggi, Marco Marano, and Nicola Pirozzi. Relationship between global end-diastolic volume and cardiac output in critically ill infants and children. *Critical care medicine*, 36(3):928–32, March 2008. ISSN 1530-0293. doi: 10.1097/CCM.0B013E31816536F7.

[16] Marko Kavcic, Brian T Fisher, Yimei Li, Alix E Seif, Kari Torp, Dana M Walker, Yuan-Shung Huang, Grace E Lee, Sarah K Tasian, Marijana Vujkovic, Rochelle Bagatell, and Richard Aplenc. Induction mortality and resource utilization in children treated for acute myeloid leukemia at free-standing pediatric hospitals in the United States. *Cancer*, 119(10):1916–23, May 2013. ISSN 1097-0142. doi: 10.1002/cncr.27957.

[17] O Sürücü, M Middeke, I Höschele, J Kalder, S Hennig, C Dietz, and I Celik. Tumour growth inhibition of human pancreatic cancer xenografts in SCID mice by cimetidine. *Inflammation research : official journal of the European Histamine Research Society ... [et al.]*, 53 Suppl 1:S39–40, March 2004. ISSN 1023-3830. doi: 10.1007/s00011-003-0318-1.

[18] Yisheng Zheng, Meng Xu, Xiao Li, Jinpeng Jia, Kexing Fan, and Guoxiang Lai. Cimetidine suppresses lung tumor growth in mice through proapoptosis of myeloid-derived suppressor cells. *Molecular immunology*, 54(1): 74–83, May 2013. ISSN 1872-9142. doi: 10.1016/j.molimm.2012.10.035.

[19] Martina Kubecova, Katarina Kolostova, Daniela Pinterova, Grzegorz Kacprzak, and Vladimir Bobek. Cimetidine: an anticancer drug? *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*, 42(5):439–44, April 2011. ISSN 1879-0720. doi: 10.1016/j.ejps.2011.02.004.

Drug-Drug Interaction Data Source Survey and Linking

Serkan Ayvaz, MS[1], Qian Zhu, PhD[2], Harry Hochheiser, PhD[3], Mathias Brochhausen, PhD[4], John Horn, PharmD[5], Michel Dumontier, PhD[6], Matthias Samwald, PhD[7] Richard D. Boyce, PhD[3]

[1]Kent State University, Kent, OH; [2]Mayo Clinic, Rochester, MN; [3]University of Pittsburgh, Pittsburgh, PA; [4]University of Arkansas for Medical Sciences, Little Rock, AK; [5]University of Washington, Seattle, WA; [6]Stanford University, Palo Alto, CA; [7]Medical University of Vienna, Vienna, Austria

Abstract

As an initial step towards the goal of a common data model for potential drug-drug interactions, we surveyed the data elements provided by the publicly available sources. Our analysis found that there is very little overlap between or across publicly available resources and that the information provided is very heterogeneous.

Introduction

Health care providers often have inadequate knowledge of what drug interactions can occur, patient specific factors that can increase the risk of harm from an interaction, and how to properly manage an interaction when patient exposure cannot be avoided. As a result, many thousands of lives are negatively affected by preventable drug-drug interactions each year. Addressing these problems is urgent as the majority of United States healthcare organizations strive to include potential drug-drug interaction (PDDI) screening in their strategies to achieve effective use of electronic health records.

We propose a new PDDI knowledge representation paradigm that we hypothesize would reduce preventable medication errors by more effectively synthesizing existing available PDDI knowledge, and more rapidly producing evidence to fill in knowledge gaps. A key component of the new paradigm is the ability to connect PDDI information from multiple sources towards the goal of obtaining more complete understanding of PDDIs. Our objective was to investigate *publicly available* (i.e., non-proprietary) PDDI information sources that may be linkable and evaluate their information coverage. We also sought to survey the data elements provided by each source as a first step toward a common data model for representing PDDIs. Our motivation for focusing on non-proprietary sources was that the number of such sources has grown in recent years and the PDDIs they provide might enhance other widely used public information systems such as Wikidata

Methods

We conducted a search of Google, PUBMED, and Embase to identify sources of PDDI information. The reference lists of relevant articles retrieved by this search were scanned for additional sources. This search was supplemented by a scan of resources provided by the Bioportal, OntoBee, and datahub for drug interaction data sets.
We downloaded public PDDI datasets identified from the aforementioned search that were available in file format or via an API. We then developed a simple PDDI data model (as a Python dictionary) that combined the data elements provided from each source. Custom Python scripts were used to translate the PDDIs listed in each source to the model. The proportion of PDDIs common between and across the downloaded datasets was examined. To enable cross-dataset comparisons, drug identifiers in each dataset were mapped to DrugBank identifiers wherever possible. For datasets where this was not the case, custom mappings were generated by finding "hub" resources on the Semantic Web that enabled a mapping from the PDDI dataset to DrugBank.

Results and Discussion

Our analysis found that that there is very little overlap between or across publicly available PDDI resources and that the information each source provides is very heterogeneous. In spite of this, our results suggest that making the sources interoperable will indeed enable a better synthesis of PDDI knowledge and making it easier to identify gaps that can be directly investigated using pharmacoepidemiology. Moreover, combing the information available across the multiple sources into the simple PDDI data model provided much richer description of these interactions. Our results indicate the importance of further research on generating high quality, complete, and consistently updated mappings between the drug terms in these information sources.

Acknowledgements

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network), the NLM (R01LM011838) and the Agency for Healthcare Research and Quality (K12HS019461).

Selective Model Averaging with Bayesian Rule Learning for Predictive Biomedicine

Jeya B. Balasubramanian, MS[1,2], Shyam Visweswaran, MD, PhD[1,2,3], Gregory F. Cooper, MD, PhD[1,2,3], Vanathi Gopalakrishnan, PhD[1,2,3]
[1]Department of Biomedical Informatics, [2]Intelligent Systems Program, [3]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA

Abstract

Accurate disease classification and biomarker discovery remain challenging tasks in biomedicine. In this paper, we develop and test a practical approach to combining evidence from multiple models when making predictions using selective Bayesian model averaging of probabilistic rules. This method is implemented within a Bayesian Rule Learning system and compared to model selection when applied to twelve biomedical datasets using the area under the ROC curve measure of performance. Cross-validation results indicate that selective Bayesian model averaging statistically significantly outperforms model selection on average in these experiments, suggesting that combining predictions from multiple models may lead to more accurate quantification of classifier uncertainty. This approach would directly impact the generation of robust predictions on unseen test data, while also increasing knowledge for biomarker discovery and mechanisms that underlie disease.

Introduction

Models that predict phenotypes and disease states from high-dimensional '-omic' datasets can lead to discovery of useful and predictive biomarkers. The typical approach for learning predictive models is to perform model selection wherein a single model is selected that summarizes the data well. However, when using real datasets there may be substantial uncertainty in choosing one model over all others, especially when the selected model is one of several models that all summarize the data more or less equally well. A sound approach in this situation is *Bayesian model averaging* (BMA) wherein the prediction for a test instance is obtained from a weighted average of the predictions of all possible models within a model space, with more probable models influencing the prediction more than less probable ones (Hoeting et al., 1999). Often in real datasets, the number of possible models is enormous, and averaging the predictions over all of them is infeasible. A practical approach is to average over a few good models, termed *selective BMA*, which serves to approximate the predictions that would be obtained from averaging over all models. The method that we describe in this paper performs selective BMA over a set of probabilistic rules as an approximation to complete BMA over all such rules.

In this paper, we extend a novel rule generation method called Bayesian Rule Learning (BRL), which identifies a single set of probabilistic classification rules learned from a training data set that can be applied to predict the class value on unseen test data. We perform selective BMA over the rule sets of BRL in order to account for model uncertainty. We compare this selective BMA approach to a model selection approach and report experimental results obtained from a range of biomedical datasets.

Background

In this section, we provide details of constrained Bayesian networks, the Bayesian scoring of models, the implementation of model selection in BRL, and the selective model averaging version of BRL (SMA-BRL).

Bayesian networks: A Bayesian network (BN) is a probabilistic graphical model that combines a graphical representation of the probabilistic dependencies between variables and the probabilistic parameters of the BN. The graphical structure is a directed acyclic graph (DAG), where the nodes represent the predictive variables and edges represent a (conditional) probabilistic dependency between corresponding variables. Absence of an edge indicates (conditional) probabilistic independence between the corresponding variables. The probabilistic parameters represent joint probability distributions over a set of predictive variables. BRL uses a Bayesian score (described below) to evaluate constrained BN structures (see Figure 1a). A complete decision tree (see Figure 1b) represents the parameters of the target node.

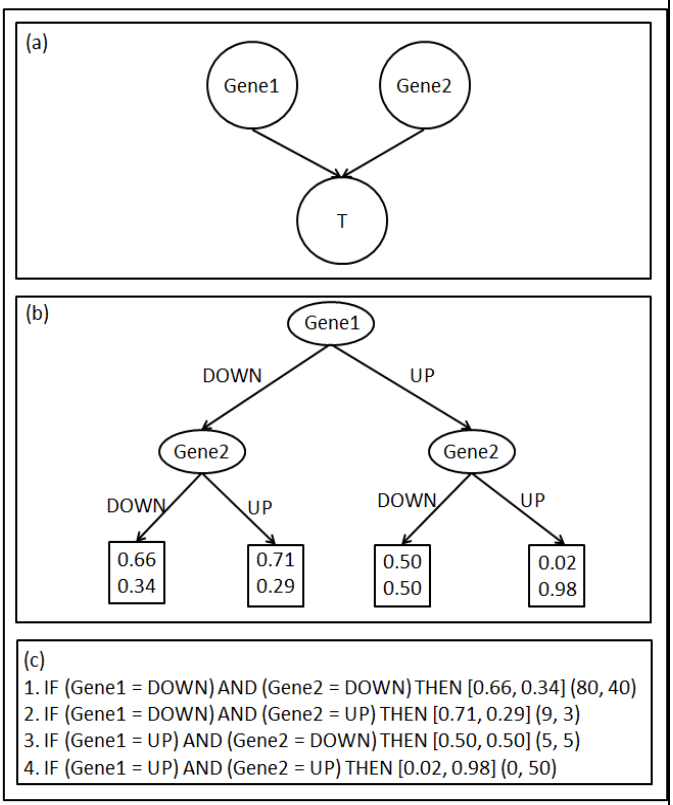


Figure 1: An example of a BN structure learned from BRL. Panel (a) displays a constrained BN structure (*S*) with two predictive variables, 'Gene1' and 'Gene2', as parents of the target variable *T*. The two predictive variables are binary with values of 'UP' and 'DOWN'. The target variable is binary having values 'Case' and 'Control'. Panel (b) shows the parameters for the target node as a complete decision tree. The interior nodes of the tree are the predictive variables (represented by ellipses) and the leaf nodes (represented by rectangles) show the probability distribution over *T*. Panel (c) shows the rule set inferred from the decision tree by BRL. Each rule antecedent is a path from a leaf to the root node. The consequent is the probability distribution of *T*. The following parentheses show the number of 'Case' instances and the number of 'Control' instances that match the antecedent, respectively.

This tree contains internal nodes that represent predictive variables and terminal nodes (leaves) that store the probability distribution over the target variable. Each leaf has a unique path to the root node. Each path represents a unique configuration of the parental states. Together, the leaves represent every possible parental state. BRL infers a set of rules (Figure 1c) from the decision tree. The rule set is the classifier model that describes the learned graphical structure and the probabilistic parameters. These rules are used to predict the target value for an unseen instance.

Bayesian score: BRL (Gopalakrishnan et al., 2010) learns a constrained BN structure (where a subset of the predictor variables have edges to the target) and evaluates it using a Bayesian score, which is proportional to the likelihood of the BN structure given the data. In this paper, we use the BDeu score (Heckerman, et al., 1995) to evaluate the BN structures. Equation 1 gives the BDeu score for the target node in the BN structure. Here, the symbol Γ represents the gamma function; j iterates through each of the q joint parental states of the target node in the BN rule-structure S; k iterates through each of the r states of the target node. N_{jk} is the number of instances (samples) in the dataset D in which the target has state k and parents of the target have state j. Here, $N_j = \sum_{k=1}^{r} N_{jk}$. The term α_0 is a user-defined parameter, which is called the *prior equivalent sample size* (*pess*). In this paper we set $\alpha_0 = 1$.

$$P(D|S) = \prod_{j=1}^{q}\left(\frac{\Gamma\left(\frac{\alpha_0}{q}\right)}{\Gamma\left(N_j+\frac{\alpha_0}{q}\right)}\cdot\prod_{k=1}^{r}\frac{\Gamma\left(N_{jk}+\frac{\alpha_0}{q\cdot r}\right)}{\Gamma\left(\frac{\alpha_0}{q\cdot r}\right)}\right). \qquad (1)$$

Methods

Algorithms: We re-implemented the beam search in the BRL, as shown in Figure 2. The beam is a priority queue of size W, which holds a set of BN structures ordered by the Bayesian score. The new implementation removes certain constraints imposed by the previous implementation, such that, we now search a larger space of models and we ensure that the final beam (of size W) returns each of the best W structures (according to the Bayesian score) evaluated by the search procedure.

Model selection in BRL returns a single model, S, from a total of W models that are generated from the training data, D, by the use of beam search. For a given vector of predictor variable values X, model S generates the posterior distribution of the target values $P(T \mid X, S)$. This does not account for the uncertainty of model S which is described by its posterior probability $P(S \mid D)$. In model selection, the selected model is assumed to have a posterior probability of 1. In reality, we are not certain that the model with the highest Bayesian score is indeed the data-generating model. Ideally, we should account for this uncertainty. In Bayesian model averaging, the predicted posterior distribution of the target is weighted by the uncertainty of the model, for all models in the model space. These terms are then summed to obtain the model averaged posterior distribution of the target. The cardinality of the

model space in BRL is $\sum_{b=0}^{B} \binom{n}{b}$, where n is the number of predictor variables and B is the maximum number of parents the target node can have. The total number of models grows rapidly in n. It is generally not feasible to enumerate every possible model. Instead, we make use of the W models already available from the existing beam search in the BRL. For model averaging, we average over these W models. This is called selective Bayesian model averaging. Equation 2 gives the average of the posterior distributions of the target node T, averaged over W models.

$$P(T|X) = \sum_{i=1}^{W} P(T|X, S_i) \cdot P(S_i|D) \quad (2)$$

The posterior probability of each model in Equation 2 is derived using Equation 3.

$$P(S_i|D) = \frac{P(D|S_i) \cdot P(S_i)}{\sum_{j=1}^{W} P(D|S_j) \cdot P(S_j)} \quad (3)$$

INPUT: A training dataset D with m instances, a set of n discrete predictor variables $\mathbf{X} = \{X_1, X_2 \ldots X_n\}$, and a discrete target variable T. The maximum number of parents that the target node T can have is *max_parents*. The maximum number of rule structures that the beam holds is W.

OUTPUT: Returns the posterior distribution of the target node given an input vector of n discrete variables.

DEFINITIONS:
T is the target node;
$Score(S) = P(D \mid S)$. This is the Bayesian score for rule structure S generated from dataset D (see Equation 1);
Q = Priority queue defined by the set $\{S_1, S_2 \ldots S_W\}$ where $Score(S_i) > Score(S_j)$, when $i < j$;
F = Priority queue of models;
V = Set of all variables in dataset D, except target variable T;
$\pi(S)$ denotes the parents in rule structure S;
max_parents = 8 (default);
W = 1000 (default);

ALGORITHM:
BRL_BeamSearch:

```
1.      Create structure S containing just target node T and place S onto Q
2.      WHILE (Q is not empty) DO:
3.          S_curr ← remove(Q)
4.          IF(F does not contain S_curr):
                place S_curr onto F
            END-IF
4.          V' = V - π(S_curr) //V_i is a set of all variables not in S_curr.
5.          IF ( (V' ≠ ϕ) AND (|π(S_curr)| < max_parents ) THEN:
6.              FOR-EACH (v in V') DO:
                    S_new ← Add v as a parent of T in S_curr.
                    IF(Q does not contain S_new):
                        place S_new onto Q
                    END-IF
                    IF(F does not contain S_new):
                        place S_new onto F
                    END-IF
                END-FOR-EACH //Ends all specialization
8.              Trim Q to the first W elements
                Trim F to the first W elements
            END-IF
        END-WHILE       //End of beam search
9.      Return F
```

(a) BRL_ModelSelection:

```
1.      F = BRL_BeamSearch:
2.      S_best ← remove(F)
3.      S_best is used to predict T.
```

(b) BRL_SelectiveModelAveraging:

```
1.      F = BRL_BeamSearch:
2.      The W models in F are used to predict T (see Equation 2).
```

Figure 2: Algorithm for model selection and model averaging in BRL.

Biomedical datasets: We analyzed the performance of SMA-BRL and BRL on 12 publicly available biomedical datasets that are listed in Table 1. It has been shown that irrelevant variables tend to introduce noise during the model search process when there are high-dimensional biomedical data with a large number of predictor variables, but relatively few samples (Liu, et al. 2012). As an initial step, we therefore applied the Partitioning-based Adaptive Irrelevant Feature Eliminator (PAIFE) to remove irrelevant features. PAIFE deems a variable as 'unconditionally relevant' by using a univariate analysis that adaptively employs the chi-square test or the Fisher's exact test. PAIFE also detects 'conditionally relevant' variables from subsets of variables, where the relationship of the variable to the target variable is conditional over other variables. The variables that are neither conditionally nor unconditionally relevant were considered irrelevant and were removed from the dataset.

Table 1. The 12 biomedical datasets used for analysis. The first eleven are genomic and the twelfth one is proteomic. The data are identified with the 'Dataset ID'. The column 'P/D' describes the type of data as Prognostic (P) or Diagnostic (D). The '# V' column is the number of predictor variables originally in the dataset. The '$\#V_{PAIFE}$' column shows the number of variables selected by PAIFE. The 'Sample Class Distribution' shows the number of samples in each class in the dataset. The 'Reference' points to the relevant literature for the dataset.

Dataset	P/D	#V	$\#V_{PAIFE}$	Sample class distribution	Reference
1	D	6584	1972	40:21:00	(Alon, et al., 1999)
2	D	12582	2371	28:24:20	(Armstrong, et al., 2002)
3	P	5372	858	69:17:00	(Beer, et al., 2002)
4	D	7129	2288	47:25:00	(Golub, et al., 1999)
5	D	7464	1880	18:18	(Hedenfalk, et al., 2001)
6	P	7129	699	40:20:00	(Iizuka, et al., 2003)
7	D	2308	832	29:25:17:12	(Khan, et al., 2001)
8	D	7129	1927	58:19:00	(Shipp, et al., 2002)
9	D	10510	6713	52:50:00	(Singh, et al., 2002)
10	P	24481	4251	44:34:00	(Veer, et al., 2002)
11	D	7039	1230	35:04:00	(Welsch, et al., 2001)
12	D	70	15	139:66	(Bigbee, et al., 2012)

Experimental methods: We wanted to evaluate and compare the predictive performance of BRL and SMA-BRL, to quantify the change in predictive performance due to model averaging. We evaluated the two algorithms, over the 12 publicly available datasets, using 10 runs of 10-fold stratified cross-validation. For a given run, the mean performance (see below) over the 10 folds was derived. We used the average of those means as an estimate of the predictive performance of the algorithm for a given dataset.

Discretization: BRL and consequently SMA-BRL require discrete values for all the variables in the input dataset. The datasets that we analyzed (see Table 1) have continuous valued predictor variables, and a discrete target variable. In the 10 runs of 10-fold cross-validation, each fold was discretized using the efficient Bayesian discretization (EBD) method (Lustgarten, et al., 2011). EBD takes a parameter λ, which determines the expected number of cut-points for each variable. For our analysis, we set $\lambda = 0.5$.

Performance measure: The performance of the algorithms was evaluated using the percentage of the area under the ROC curve (AUC). The area under the ROC curve is typically used as a summary statistic of discrimination. The AUC is equivalent to the probability that a randomly chosen case from the negative class will have a smaller predicted probability of belonging to the positive class than a randomly chosen case from the positive class.

The average AUCs obtained from the two algorithms for each of the 12 datasets, over 10 runs of 10-fold stratified cross-validation, is analyzed using two statistical tests. We used the tests to check whether the difference between the performances of the two classifiers over the 12 datasets is non-random. The tests included (1) significance testing with the Wilcoxon paired-samples two-sided signed ranks test, and (2) effect size testing with paired-samples two-tailed t-test. We used the Statistics Toolbox from MATLAB to perform these tests (MATLAB and Statistics Toolbox Release 2013b, The MathWorks, Inc., Natick, Massachusetts, United States).

Results and discussion

The average AUCs obtained from the two algorithms for each of the 12 datasets is shown in Table 2. The result from the significance tests show that SMA-BRL is statistically significantly better than BRL based on these AUC values.

Table 2. Average AUCs obtained from BRL and SMA-BRL using 10 runs of 10-fold cross-validation for the 12 datasets described in Table 1. For each dataset, the result of the better performing algorithm is shown in bold. The last row shows the average from the 12 datasets and the standard error of mean (SEM).

Dataset	BRL	SMA-BRL
1	**99.50**	**99.50**
2	95.12	**95.67**
3	60.14	**60.25**
4	91.88	**93.82**
5	94.13	**100.00**
6	57.19	**58.13**
7	84.67	**86.55**
8	81.58	**82.87**
9	90.87	**90.95**
10	86.12	**86.50**
11	95.42	**97.92**
12	80.96	**82.28**
Average ± SEM	84.80 ± 3.90	86.20 ± 4.05

The non-parametric Wilcoxon paired-samples signed ranks test, with significance level $\alpha = 0.05$, shows that SMA-BRL performs statistically significantly better than BRL with a p-value of 9.765×10^{-4}. The paired-samples two-tailed t-test, with significance level $\alpha = 0.05$, also shows that SMA-BRL performs statistically significantly better than BRL with a p-value of 0.0122. The 95% confidence interval of the mean of the difference between the column values of BRL and SMA-BRL in Table 2, based upon the t-distribution is [-2.438, -0.372].

We observe that the difference between the average AUC, for BRL and SMA-BRL, across the 12 datasets is small. We also observe that for each of the 12 datasets we analyzed in this paper, SMA-BRL either obtains an equivalent or better average AUC performance than BRL. Note that the SMA-BRL uses the same search engine as the BRL. The BRL generates *W* models but only one is selected and used for inference on a test case. SMA-BRL makes use of all the *W* models for its inference. Therefore, SMA-BRL only requires an additional constant time operation during the model inference step. As a result, SMA-BRL returns a robust classifier, which is worth exploring, at very little additional computational cost. A limitation of SMA-BRL when compared to BRL is that the predictions based on SMA-BRL involve the weighted inference of *W* probabilistic rules, which is more complex to understand than the inference of a single rule in BRL.

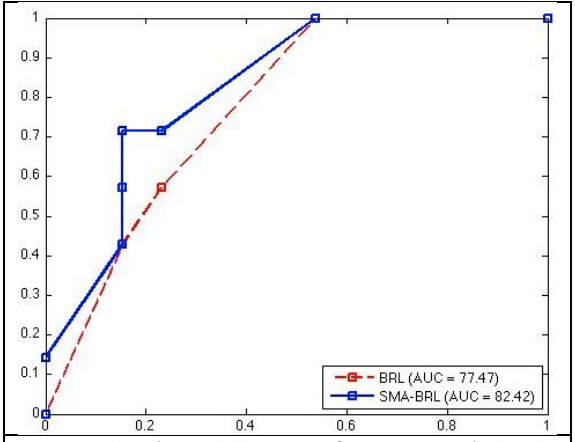


Figure 3: The ROC curve for BRL and SMA-BRL, on one training fold of dataset 12 (Table 1).

Case Study: We examined the models learned by BRL and SMA-BRL on one of the training folds of dataset 12 (see Table 1). The ROC curve of the models is shown in Figure 3. The AUC of the BRL model is 77.47 and of the SMA-BRL model is 82.42. The BRL model included three biomarkers (MIF, Thrombos, and SAA) and the SMA-BRL model, in addition to the three biomarkers, included five more biomarkers (IL-8, IGFBP-1, PROLACTI, TTR, and RANTES). In future work, we plan to study the relative importance of these variables and their biological significance.

Conclusion

SMA-BRL accounts for model uncertainty, which the model selection method BRL ignores. SMA-BRL generates robust predictions from a committee of plausible models. The strong theory underlying Bayesian model averaging is supported by the results from our analysis of 12 datasets. Moreover, since SMA-BRL only averaged over the models encountered in the BRL search, the computational time complexity of the two algorithms is almost identical. Thus, the improved results achieved with SMA-BRL are obtained essentially for free. Overall, these results support using model averaging when predicting outcomes in biomedical datasets that are similar to the 12 datasets analyzed in this paper.

Acknowledgements

The authors thank the anonymous reviewers for their insightful comments that helped tailor the paper for the intended audience. The authors gratefully acknowledge grant number R01-LM010950 from the National Library of Medicine. VG was funded in part by grants R01GM100387 and P50CA090440 from the National Institutes of Health. GFC was funded in part by NIH grant R01LM010020 and NSF grant IIS0911032. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382-401.
2. Gopalakrishnan, V., Lustgarten, J. L., Visweswaran, S., & Cooper, G. F. (2010). Bayesian rule learning for biomedical data mining. *Bioinformatics*, 26(5), 668-675.
3. Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning, 20(3), 197-243.
4. Lustgarten, J. L., Visweswaran, S., Gopalakrishnan, V., & Cooper, G. F. (2011). Application of an efficient Bayesian discretization method to biomedical data. *BMC Bioinformatics*, 12(1), 309.
5. Liu, G., Kong, L., & Gopalakrishnan, V. (2012). A Partitioning Based Adaptive Method for Robust Removal of Irrelevant Features from High-dimensional Biomedical Datasets. *Proceedings of the AMIA Summits on Translational Science*, 2012, 52.
6. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12), 6745-6750.
7. Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., Den Boer, M. L., Minden, M. D., … Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30, 41–47.
8. Beer, D. G., Kardia, S. L. R., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., … Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8, 816–824.
9. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.
10. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., … Sauter, G. (2001). Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, 344, 1–6.
11. Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., … Hamamoto, Y. (2003). Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, 361, 923–929.
12. Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., … Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, *7*, 673–679.
13. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., … Golub, T. R. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8, 68–74.
14. Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., … Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1, 203–209.
15. Veer, L. van't, Dai, H., & Vijver, M. Van De. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*.
16. Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., ... & Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Sciences,* 98(3), 1176-1181.
17. Bigbee, W. L., Gopalakrishnan, V., Weissfeld, J. L., Wilson, D. O., Dacic, S., Lokshin, A. E., & Siegfried, J. M. (2012). A multiplexed serum biomarker immunoassay panel discriminates clinical lung cancer patients from high-risk individuals found to be cancer-free by CT screening. *Journal of Thoracic Oncology*, 7(4), 698.

Generalized Linear Models for Identifying Predictors of the Evolutionary Diffusion of Viruses

Rachel Beard[1], Daniel Magee[1,], Marc A. Suchard MD, PhD[2], Philippe Lemey PhD[3], Matthew Scotch PhD, MPH[1]
[1]Arizona State University, Tempe, AZ, USA
[2]University of California, Los Angeles, CA, USA
[3]KU Leuven, Leuven, Belgium

Abstract

Bioinformatics and phylogeography models use viral sequence data to analyze spread of epidemics and pandemics. However, few of these models have included analytical methods for testing whether certain predictors such as population density, rates of disease migration, and climate are drivers of spatial spread. Understanding the specific factors that drive spatial diffusion of viruses is critical for targeting public health interventions and curbing spread. In this paper we describe the application and evaluation of a model that integrates demographic and environmental predictors with molecular sequence data. The approach parameterizes evolutionary spread of RNA viruses as a generalized linear model (GLM) within a Bayesian inference framework using Markov chain Monte Carlo (MCMC). We evaluate this approach by reconstructing the spread of H5N1 in Egypt while assessing the impact of individual predictors on evolutionary diffusion of the virus.

Introduction

Bioinformatics and phylogeography models use viral sequence data to analyze spread of epidemics and pandemics. However, few of these models have included analytical methods for testing whether certain predictors such as population density, rates of disease migration, and climate are drivers of spatial spread. While spatial epidemiology has successfully developed models of environmental predictors such as global mobility and air travel, these models remain disconnected to molecular sequence data that are analyzed through bioinformatics and phylogeography applications to unlock information about virus coalescence, spatial spread, and gene flow.[1] Combining spatial epidemiology and molecular sequence data can lead to discoveries about risk of transmission between animals and humans as well as the relationship between geography and genetic evolution of the virus. In addition, understanding the specific factors that influence spatial diffusion of viruses is critical for targeting public health interventions and limiting spread. In this study, we describe the application and evaluation of a phylogeographic model that integrates demographic and environmental factors. Here we focus on a variant clade of H5N1 viruses in Egypt and its countrywide diffusion among avian and human hosts. This approach is generalizable to other RNA viruses and may enhance both public health prevention and response by identifying the drivers that are most vital to viral spread.

Background

Many emerging or re-emerging infectious diseases are zoonotic in origin, and pose significant threats to human and animal health.[2] There are many potential drivers of transmission between animals and humans and many of these drivers likely vary between countries. This variation could be caused by climate differences, population sizes, and living conditions, as well as cultural practices related to food preparation and distribution. In response to these complexities, many epidemiologic models have studied potential contributors such as human and avian population densities, or precipitation.[3] For example Van Boeckel *et al.* examined anthropogenic and ecological variables relating to avian species within developed regions in Asian farming communities following flood conditions,[4] while Tamerius *et al.* observed the effects of temperature, humidity, and precipitation on H5N1 spread in tropical climates.[5] While this research has resulted in valuable epidemiologic insights, it has traditionally ignored the information about the evolutionary processes occurring within the viral genome. Phylodynamic analysis of RNA viruses can lead to crucial information regarding transmission, genetic diversity and selection, as well as epidemiologic characteristics.[6] Bioinformatics and phylogeography techniques have enabled researchers to depict local and global virus spread, providing valuable information to the public health community as to the origin and epidemic patterns of spread. For instance, Lam *et al.* determined that the spread of influenza A subtype H5N1 was likely introduced into Indonesia by a single introduction in East Java in approximately 2002, followed by both an east and westward migration throughout the country.[7] Bioinformatics approaches such as these are informative; though few incorporate demographic and environmental factors often used in epidemiology. Ypma *et al.* demonstrate this concept by including geographic and temporal elements as well as genetic data to estimate the

migration patterns of influenza A subtype H7N7 in the Netherlands.[8] By taking an integrated approach, this work highlighted the estimates of certain drivers on evolutionary transmission with greater accuracy.[8] The same group also demonstrated that using within-host dynamics and genetic data of pathogens to simultaneously generate both the phylogenetic tree and transmission route leads to more accurate models and plausible estimation of connecting variables.[9] Thus, epidemiologic and viral phylogenetic approaches have been incorporated into a rough framework which join evolutionary and ecologic dynamics to explain spatial diffusion.[10] Phylogeography naturally compliments models based on observed epidemiologic data, as the genomic data can provide a record by which to confirm or reject hypothesized patterns of viral spread. Our aim is to demonstrate the utility of combining epidemiologic and phylogeographic approaches to identify drivers of virus diffusion. We evaluate this approach by reconstructing the spread of H5N1 in Egypt while assessing the impact of individual predictors on evolutionary diffusion of the virus.

Methods

A Bayesian generalized linear model (GLM) approach was adopted which was developed by Lemey *et al.*, in which the spatiotemporal patterns of viral diffusion are reconstructed while potential contributing factors are simultaneously assessed.[11] We use the work of Scotch *et al.*[12] as a basis by which to analyze the potential environmental drivers of highly pathogenic avian influenza (HPAI) H5N1 movement among multiple hosts by considering discrete geographic locations within Egypt. We chose to focus on Egypt because it has recently emerged as an epicenter for H5N1, with 173 human cases reported to the World Health Organization (WHO) as of June 2013.[12] In addition, the local cultures prefer to obtain their poultry via live bird markets which create an atmosphere of high human-avian transmissibility.

Sequence data

We used the same dataset described by Scotch *et al.*[12] that included 226 H5N1 hemagglutinin (HA) sequences previously isolated, however we excluded two sequences for which the host was recorded as environmental. Sequences collected from avian (n=210) and human (n=14) hosts in Egypt spanning 2007-2012. The sequences were selected based on their Egyptian origin and classification within the recently defined variant subclade 2.2.1.1. published by WHO.[13]

We reconstructed the spread of H5N1 in Egypt using a discrete phylogeography approach while estimating the effect of a diverse set of variables on phylogeographic diffusion within a GLM. This process was implemented using the development version of the BEAST software package, available at http://code.google.com/p/beast-mcmc/, which uses a Bayesian Markov Chain Monte Carlo (MCMC) analysis.[14] We modeled sequence evolution using the generalized time-reversible (GTR) model of nucleotide substitution, while using a relaxed molecular clock. Multiple chain lengths were tested using Tracer,[15] with the final run set at 20 million.

Generalized linear model

We tested the effect of predictors on spatial spread while reconstructing the spatiotemporal history. Here, we used modeling techniques described in Lemey *et al.*,[11] and innovative methods for Bayesian phylogeographic inference of phylogenetic history and discretized diffusion processes.[16] We utilized a GLM model by integrating diffusion of viral spread as a non-reversible continuous time Markov chain processes expressed as a K x K infinitesimal rate matrix of location change (Λ) among K discrete locations.[11] We represented all rates of movement Λ_{ij} using a log linear function to incorporate a set of n predictors on the log-scale.

$$\log\Lambda_{ij} = \beta_1\delta_1\log(p_1) + \beta_2\delta_2\log(p_2) + \ldots + \beta_n\delta_n\log(p_n)$$ [11]

Here, β signifies the contribution of a given predictor to the model, and δ is a binary indicator (0, 1) variable that oversees whether a particular predictor is to be incorporated in the model.[17] This allows for Bayesian stochastic search variable selection (BSSVS),[16-18] in which posterior probabilities of all possible models that may or may not include a given predictor are estimated, as discussed in Lemey *et al,* 2009.[17, 18] We utilized a Bernoulli prior probability distribution for δ as in Lemey *et al.* 2012, to place equal probability of inclusion or exclusion of predictors.[11]

We selected local predictors based on feedback from experts who study H5N1 in Egypt.[19] These predictors were

chosen to represent genomic, geographical, demographical, and numerical indicators to develop a preliminary model and include:

***Avian and human population density**:* We incorporated population density for all possible origins and destinations for both humans and chickens from City Population, an online resource for worldwide population statistics, and the Food and Agriculture Organization of the United Nations (FAO).[20, 21]

Latitude: We obtained the latitude of the centroid location for each governorate in order to reflect diverse climatic conditions within the country by using GeoNames.[22] While this likely does not reflect the true locations of where sequences were collected, this method was adopted to impose uniformity across the model.

Distance: We calculated the distance between governorates using the centroid latitude and longitude obtained from GeoNames.[22]

Case and Sequence counts: We obtained estimates of human and avian H5N1 cases for each governorate from the FAO for the years of 2006-2012.[19] We averaged these to obtain the final predictor values for our model. The sequences incorporated into the phylogeographic analysis were differentiated by the location from which they were isolated for both human and avian sequences. We included these variables not to explain diffusion, but rather to minimize bias on predictors being tested by indicating the sample sizes at particular locations throughout viral spread.

We log transformed and standardized all predictors before their incorporation into the model.

Evaluation of predictor inclusion

Following Lemey *et al.*[11, 16] we determined the support for predictors within the model using Bayes factors (BFs). To calculate the BFs, the posterior odds of predictor inclusion were divided by their prior odds:

$$\text{BF} = \left(\frac{pi}{1-pi}\right)/\left(\frac{qi}{1-qi}\right) \quad [11]$$

Here p_i represents an estimate of the posterior probability that a given predictor is included while q_i represents the prior probability. For this study, the BF cutoff for support within the model was set at 3. We implemented a technique for adjusting β to a fixed correlation $X'X$ in order to account for possible high correlation between predictors. Finally, we evaluated δ under a bit flip operator as discussed by Drummond *et al*. in greater detail.[23]

Results

The BF results suggest the importance of avian populations to the viral diffusion of H5N1 clade 2.2.1.1 in Egypt (figure 1). Most notably, avian population density at the origin had a strong support for inclusion within the model of viral spread with a BF score of 22.3. Additionally, we derived the 95% Bayesian credible interval for the coefficient of each predictor which indicates the level of uncertainty of a particular variable. The inclusion of avian densities at the origin within the model was also supported in this respect, with a credible interval which did not span zero. However, the credible interval for distance, latitude of origin, and human density at the origin did span zero. Compared to avian densities at the origin, human population density did not indicate nearly the degree of support. For both populations the origin achieved a higher probability of inclusion compared to the destination of spread during the observed time period. Other predictors included in the model such as distance between the origin and destination of spread and latitude within Egypt achieved negligible BF scores and inclusion probability. Human density, avian density and latitude at the destination were not supported within model as BF values dropped to approximately 1 or below. Finally, while the variables relating to sample size of sequences and case counts do not directly contribute to the model, their inclusion lends increased credibility for the predictors relating to the avian host data, in particular the avian sequence data which received a BF score of 61.5 and variables associated with the human host obtained unsupportive BF values.

Discussion

Mitigation and prevention of infectious disease is essential to population health, and to achieve these goals we must first understand the processes that drive the spread of viruses such as influenza. Our preliminary work indicates the potential to uncover variables of interest for a particular virus and region, which highlight the integration of

epidemiologic and phylogenetic approaches. Of the tested predictors for H5N1 spread within Egypt, we have found host population densities within the region to be strong indicators for viral dispersal and highly supported for inclusion within the model by BF values. These results are consistent with the nature of close proximity within large populations, and with other findings related to H5N1 risk factors. For instance, Martin *et al*. found that chicken and

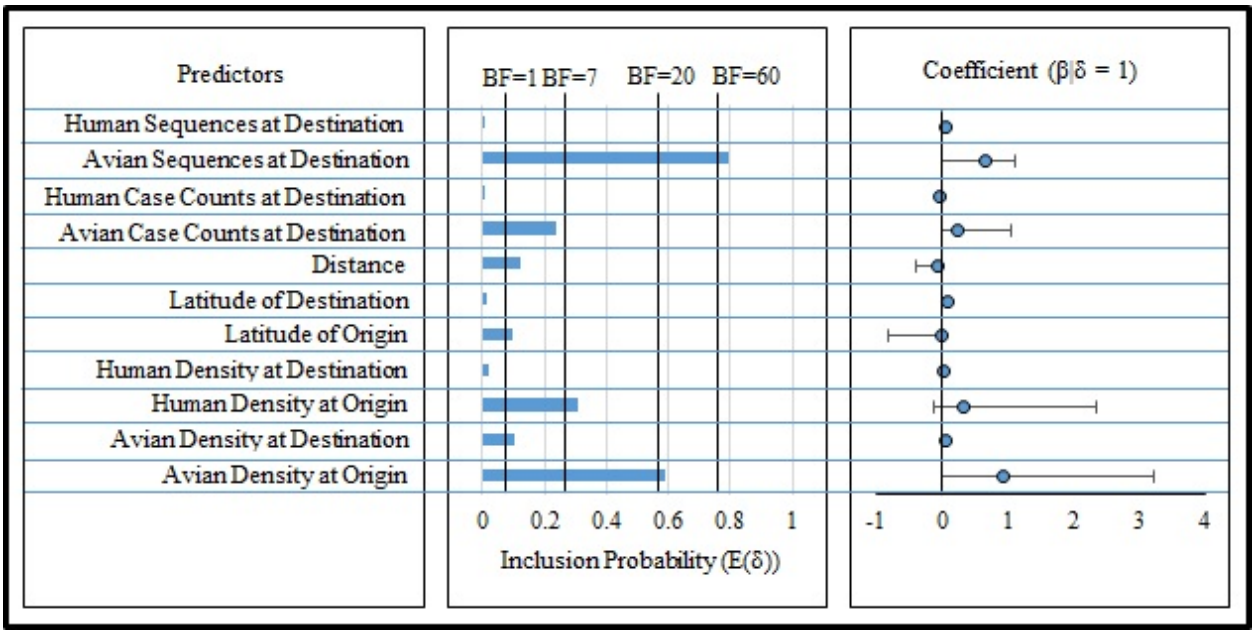


Figure 1. Predictors of H5N1 diffusion in Egypt. Inclusion probability defined by indicator expectations E(δ), which reflects the likelihood of meaningful impact of the predictor on viral diffusion. Bayes Factor (BF) support values shown at the top of the figure and are indicated by vertical lines. Coefficient (β|δ=1) represents the contribution of each predictor, with the 95% credible interval represented by brackets.

human density in China was a leading contributor to risk of infection.[24] However, we do not preclude the possibility of other potential underlying dynamics driving influenza H5N1 in Egypt. While our case study involved influenza, this approach can be applied to other RNA viruses as they have shorter genomes and more rapid nucleotide substitutions compared to other pathogens.[25]

Limitations

There are several limitations of this work, largely related to incomplete or outdated data sources. Our assignment of the centroid of each governorate as the latitude for discrete locations can only approximate the geographic distribution of viral spread. In addition, it is nearly certain the actual number of case counts observed in human and avian populations was not represented as mild cases may go unrecognized. Case counts can also vary year-to-year, possibly indicating the influence of another predictor. This possibility is overlooked using our current method of averaging a range of years. Additional sequencing of collected viruses from known cases would also aid our depiction of the spatial distribution, particularly human sequences as this data is sparse. Finally, estimates of avian population densities used here were collected in 2005, which may over or underestimate actual densities throughout our study period.

Conclusion

We demonstrate the potential of phylogeography and bioinformatics techniques to incorporate traditional epidemiologic data for understanding the evolutionary diffusion of viruses. Future work will involve testing additional variables that are indicated in viral proliferation within Egypt. Predictors of interest include domestic avian population ranges with migratory bird habitat overlap, cross species spill over migration rates, as well as the recent discovery of an important shift in amino acid composition of the hemagglutinin cleavage site to viral pathogenicity within Egyptian strains.[26]

Acknowledgments

The project described was supported by award number R00LM009825 from the National Library of Medicine to MS and by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant Agreement no. 278433-PREDEMICS and ERC Grant agreement no. 260864 to PL. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health. The authors would like to thank the Arizona State University Advanced Computing Center (A2C2) for the use of the Saguaro supercomputer.

References

1 Viboud C, Bjørnstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. Science. 2006 April 21, 2006;**312**(5772):447-51.
2 Krauss H. Zoonoses: Infectious Diseases Transmissible from Animals to Humans: ASM Press; 2003.
3 Herrick K, Huettmann F, Lindgren M. A global model of avian influenza prediction in wild birds: the importance of northern regions. Veterinary Research. 2013;**44**(1):42.
4 Van Boeckel TP, Thanapongtharm W, Robinson T, Biradar CM, Xiao X, Gilbert M. Improving Risk Models for Avian Influenza: The Role of Intensive Poultry Farming and Flooded Land during the 2004 Thailand Epidemic. PloS one. 2012;**7**(11):e49528.
5 Tamerius JD, Shaman J, Alonso WJ, et al. Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates. PLoS pathogens. 2013;**9**(3):e1003194.
6 Chu P-Y, Ke G-M, Chen P-C, Liu L-T, Tsai Y-C, Tsai J-J. Spatiotemporal Dynamics and Epistatic Interaction Sites in Dengue Virus Type 1: A Comprehensive Sequence-Based Analysis. PloS one. 2013;**8**(9):e74165.
7 Lam TT-Y, Hon C-C, Lemey P, et al. Phylodynamics of H5N1 avian influenza virus in Indonesia. Molecular ecology. 2012;**21**(12):3062-77.
8 Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proceedings of the Royal Society B: Biological Sciences. 2012 February 7, 2012;**279**(1728):444-50.
9 Ypma RJF, van Ballegooijen WM, Wallinga J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. Genetics. 2013 September 13, 2013.
10 Grenfell BT, Pybus OG, Gog JR, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science. 2004 January 16, 2004;**303**(5656):327-32.
11 Lemey P, Rambaut A, Bedford T, et al. The seasonal flight of influenza: a unified framework for spatiotemporal hypothesis testing. arXiv:12105877v1. 2012.
12 Scotch M, Mei C, Makoynnen Y, et al. Phylogeography of Influenza A H5N1 Clade 2.2.1.1 in Egypt. Unpublished. 2013.
13 WHO. H5N1 avian influenza: Timeline of major events. 2012.
14 Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012 Aug;**29**(8):1969-73.
15 Rambaut A. Tracer v1.5 [Internet]. c2009. [updated 2009 Nov 30; cited 2013 Sep 26] Available from http://treebioedacuk/software/tracer/
16 Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol. 2009;**5**(9):e1000520.
17 Kuo L, Mallick B. Variable Selection for Regression Models. Sankhya. 1998;**60**(1):65-81.
18 Chipman H, George E, McCulloch R. BART: Bayesian Additive Regression Trees. The Annals of Applied Statistics. 2010;**4**(1):266-98.
19 Arafa A. Egypt Clade 2.2.1.1. [online]. E-mail to Matthew Scotch (matthew.scotch@asu.edu). 2013 Aug 12 [cited 2013 Sep 30].
20 FAO. Animal Production and Health Division [Internet]. Global Livestock Production and Health Atlas. c2011. [updated 2013 Mar; cited 2013 Sep 26] Available from http://kidsfaoorg/glipha/indexhtml

21 Egypt CAfPMaS. Arab Republic of Egypt [Internet]. c2012. [updated 2012 Jul 05; cited 2013 Sep 26]. Available from http://www.citypopulation.de/Egypt.html.
22 Geonames.org. [Internet]. Egypt. c2013. [updated 2013 Apr 30; cited 2013 Sep 26] Available from http://wwwgeonamesorg/EG/administrative-division-egyptthtml
23 Drummond A, Suchard M. Bayesian random local clocks, or one rate to rule them all. BMC Biology. 2010;**8**(1):114.
24 Martin V, Pfeiffer DU, Zhou X, et al. Spatial Distribution and Risk Factors of Highly Pathogenic Avian Influenza (HPAI) H5N1 in China. PLoS pathogens. 2011;**7**(3):e1001308.
25 Holmes EC. The phylogeography of human viruses. Molecular ecology. 2004;**13**(4):745-56.
26 Yoon S-W, Kayali G, Ali MA, Webster RG, Webby RJ, Ducatez MF. A Single Amino Acid at the Hemagglutinin Cleavage Site Contributes to the Pathogenicity but Not the Transmission of Egyptian Highly Pathogenic H5N1 Influenza Virus in Chickens. Journal of Virology. 2013 April 15, 2013;**87**(8):4786-8.

Heterogeneity within and across Pediatric Pulmonary Infections: From Bipartite Networks to At-Risk Subphenotypes

Suresh K. Bhavnani, PhD[1], Bryant Dang, BS[1], Maria Caro, MS[1], Gowtham Bellala, PhD[2], Shyam Visweswaran, MD, PhD[3], Asuncion Mejias, MD PhD[4], Rohit Divekar, MBBS, PhD[5]

[1]Inst. for Translational Sciences, Inst. for Human Infections and Immunity, Univ. of Texas Medical Branch, Galveston, TX; [2]Hewlett Packard Laboratories, Palo Alto, CA; [3]Department of Biomedical Informatics, Univ. of Pittsburgh, Pittsburgh, PA; [4]Div. of Pediatric Infectious Diseases, Ohio State University, Columbus, OH; [5]Division of Allergic Diseases, Mayo Clinic, Rochester, MN

Abstract

Although influenza (flu) and respiratory syncytial virus (RSV) infections are extremely common in children under two years and resolve naturally, a subset develop severe disease resulting in hospitalization despite having no identifiable clinical risk factors. However, little is known about inherent host-specific genetic and immune mechanisms in this at-risk subpopulation. We therefore conducted a secondary analysis of statistically significant, differentially-expressed genes from a whole genome-wide case-control study of children less than two years of age hospitalized with flu or RSV, through the use of bipartite networks. The analysis revealed three clusters of cases common to both types of infection: *core cases* with high expression of genes in the network core implicated in hyperimmune responsiveness; *periphery cases* with lower expression of the same set of genes indicating medium-responsiveness; and *control-like cases* with a gene signature resembling that of the controls, indicating normal-responsiveness. These results provide testable hypotheses for at-risk subphenotypes and their respective pathophysiologies in both types of infections. We conclude by discussing alternate hypotheses for the results, and insights about how bipartite networks of gene expression across multiple phenotypes can help to identify complex patterns related to subphenotypes, with the translational goal of identifying targeted therapeutics.

Introduction

Most children by the age of two have been infected by RSV or influenza, both of which are associated with acute morbidity and mortality[1]. While many recover through a normal immune response, a subset develops severe disease requiring hospitalization. In fact, RSV is the leading cause of hospitalization for infants and young children worldwide, and is associated with long-term morbidity and risk for developing future chronic and recurrent wheezing[2]. Studies have identified epidemiological (e.g., second hand smoke, atopy[3]), and underlying medical conditions (e.g., congenital heart disease, prematurity or chronic lung disease[4]) as risk factors for flu and RSV infections associated with severe disease. Because the majority of children hospitalized with respiratory illnesses are previously healthy, researchers have hypothesized that the different clinical responses to such infections could be the result of host-specific genomic and immune responses that predispose patients to more severe disease[5].

Unfortunately, much remains to be discovered about these differentiating host-specific genomic and immune mechanisms because most of the research has either been conducted (1) *in vitro* using infection assays of human cells; (2) mouse models; or (3) used single markers and univariate methods to analyze a limited repertoire of analytes as potential biomarkers in humans. For example, the latter approach has found that certain cytokines (IL-4, IL-13, IL-10, IL-8), cytokine receptors (IL-4 receptor α, IL-8 receptor), innate receptors (TLR4), and even non-immune proteins like surfactant could be used as potential biomarkers to predict severe RSV infection[6]. While such studies have provided key insights into biomarkers activated in mice and certain patient populations, to the best of our knowledge no studies have used multivariate methods to analyze the heterogeneity of responses using whole-genome human data during naturally acquired flu or RSV infections. Such multivariate analysis could help to identify molecular biomarkers for at-risk subsets of patient, and in the case of RSV, pathways that predispose them to later chronic recurrent wheezing and asthma.

We therefore used bipartite networks to conduct a multivariate analysis of patients with either flu or RSV, with the goal of identifying subphenotypes and molecular pathways that were common to both types of infections. Such an approach could result in general approaches to identify and treat at-risk patients with either type of infection.

Methods

Our research began with the question: *How do differentially expressed genes that are common to flu and RSV infections co-express across patients with either type of infection?* To address our research question, we made critical decisions related to data selection, and data analysis as discussed below:

Data Selection. Our study was based on a secondary analysis of a publicly available dataset downloaded from the Gene Expression Omnibus (ID: 200034205). The data consisted of 28 flu and 51 RSV previously-healthy infected children less than 2 years of age, with confirmed microbiologic diagnosis of infection, and who were hospitalized due to severe illness. The data also included 22 age, gender, and race matched healthy controls. As reported in the primary study[7], peripheral blood mononuclear cells (PBMCs) of naturally-infected subjects were collected between 42 to 72 hours after hospital admission, and their disease severity score (aggregated measure of percutaneous O_2 saturation, respiratory rate, subcostal retractions, general appearance, and auscultation) was recorded on a scale of 1-15. The cellular RNA was extracted, their expression measured using the whole-genome Affymetrix HG-U133 plus 2.0 chip array, and the results adjusted using a single standard curve. Furthermore, the primary study identified statistically-significant (FDR corrected) genes in each infection, and selected 18 highly-significant, differentially-expressed genes that were common to both infections for univariate analysis, with the goal of identifying pathways associated with the top-ranked genes in both illnesses.

In contrast to the above primary analysis, our secondary analysis of the data consisted of a multivariate analysis of all 101 subjects, and the above 18 genes that were common to both types of infection.

Data Analysis. Our analysis consisted of two steps: (1) **exploratory visual analysis** to identify emergent bipartite relationships between patients and genes; and (2) **quantitative analysis** suggested by the emergent visual patterns. This two-step method was motivated by our earlier studies[8-10], which have demonstrated that bipartite networks can reveal complex patterns each prompting the use of quantitative methods that make the appropriate assumptions about the underlying data.

1. Exploratory Visual Analysis was conducted using network visualization and analysis[11]. Networks are increasingly being used to analyze a wide range of molecular phenomena such as gene and protein-protein interactions[12-13], and to assess their relationships to diseases, symptoms, and syndromes. A network consists of nodes and edges; nodes represent one or more types of entities (e.g., patients or genes), and edges between the nodes represent a specific relationship between the entities. Figure 1 shows a bipartite network[11] where edges exist only between patients and genes.

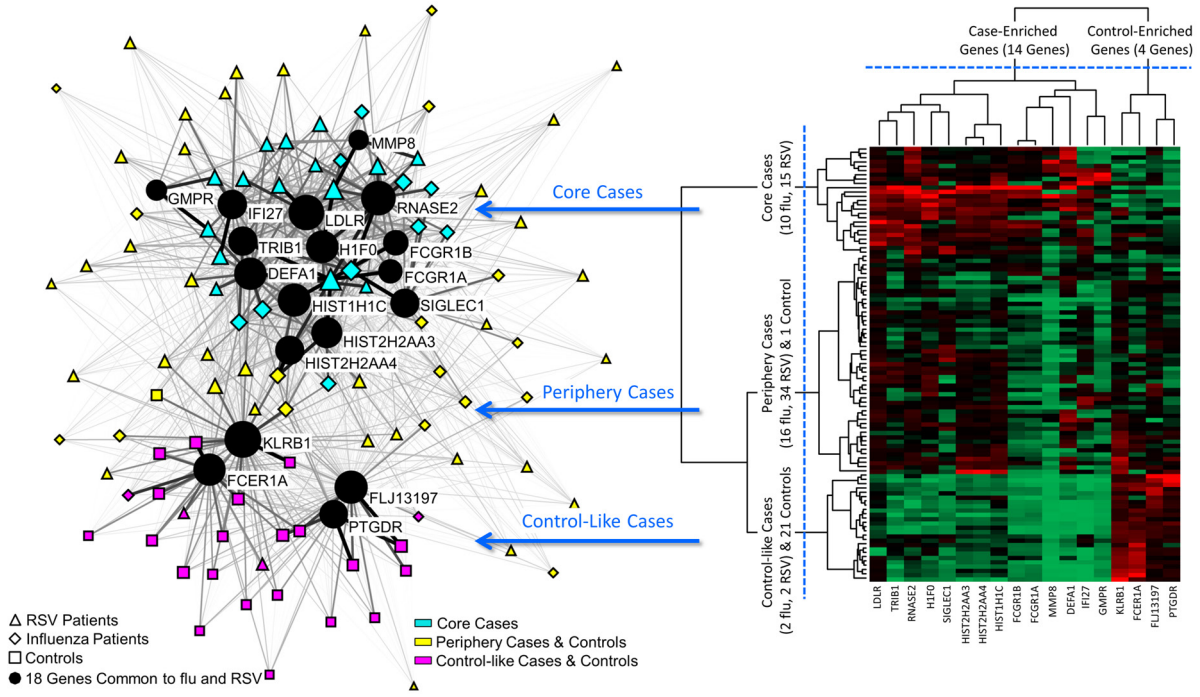
Edge weights in the network were used to represent the strength of the genes expression values for each patient-gene pair. Because the genes had different expression ranges, we used the min-max normalization method (which does a linear mapping of each genes expression to range from 0-1, and therefore preserves the relative distances between values to enable comparison). As shown in Figure 1, the edge thicknesses were drawn to be proportional to these normalized expression values. *Node diameter* was used to represent the sum of the edge weights connected to it (also referred to as the weighted degree centrality). This enabled a rapid visual inspection to determine for example, which patients have overall high aggregate expression values, and how such patients relate to the rest of the network. Finally, the *node shape* was used to represent phenotype or genes (triangles=RSV, diamonds=flu, squares=controls, circles=genes), and *node color* was used to represent members of a cluster based on hierarchical cluster analysis.

Global patterns between subjects and genes in the network were visualized and analyzed using the *Kamada-Kawai* layout algorithm[14] in Pajek (version 3.13). As shown in Figure 1, the algorithm pulls together nodes that are strongly connected, and pushes apart nodes that are not. This algorithm is fast but approximate and is well-suited for medium sized networks consisting of between 100-1000 nodes[15]. The result is that nodes with a similar pattern of connections (e.g., the gene nodes IFI27 and TRIB1 in the top of the network in Figure 1A) are placed close to each other.

A key advantage of a network representation is the simultaneous visualization of multiple **raw values** (patient-gene associations, expression values), **aggregated values** (sum of gene expression values), and **emergent global patterns** (clusters) in a *uniform* visual representation. Such a representation enables the rapid generation of hypotheses based on complex multivariate relationships, which can be verified through appropriate quantitative methods.

2. Quantitative Analysis was conducted using three measures to verify the insights derived from the exploratory visual analysis. These methods were selected based on their appropriateness to the emergent patterns in the network.

(a) ***Agglomerative Hierarchical Clustering***. Because the network layout suggested a distinct clustering combined with a *core-periphery* topology (nodes with high overall edge weights in the core, and nodes with low overall edge weights in the periphery[11]), we used the agglomerative hierarchical clustering method. The clustering was done using the Euclidean dissimilarity measure with the Ward linkage function, and the number of clusters and their



A. Bipartite Network B. Heatmap with Dendrograms

Figure 1. **A.** A bipartite network (automatically laid out by the *Kamada-Kawai* algorithm[14]) shows how 18 genes (circular nodes) co-occur across 101 subjects (triangle, diamond and square nodes). The size of the nodes is proportional to the sum of the edge weights (representing normalized expression) that connect to them, and the thickness of edges is proportional to gene expression values. Therefore subjects with high total expression have large nodes, and higher expression is represented by thicker edges. The network has an overall distinct cluster topology that separates most cases from the controls, in addition to a core-periphery topology of cases. **B.** A heatmap with dendrogram generated through hierarchical clustering helped to identify the boundaries of three subject clusters, which were superimposed onto the network using colored nodes to denote cluster membership.

boundaries were determined based on natural breaks in the patient and gene dendrograms. The cluster boundaries were superimposed onto the network by using identical node color to denote cluster membership.

(b) ***Clusteredness***. To test whether the clusters in the network could have occurred by chance, we compared the variance, skewness, and kurtosis of the dissimilarities in the data, to 1000 random permutations of this data. For each network permutation, we preserved the size of the network (number of nodes and edges), in addition to the edge weight distribution of each patient when analyzing the patient dendrogram, and the edge weight distribution for each gene when analyzing the gene dendrogram. Significant breaks in the patient, or gene dendrograms would result in a significantly larger variance, skewness, and kurtosis of the dissimilarity measures, compared to the same measures generated from the random networks.

(c) ***Relationship to Clinical Outcomes***. Because the cases had a core-periphery topology, we used the Mann Whitney *U* test to analyze whether the patients in the case core had a higher overall gene expression (using *weighted degree centrality* calculated by adding each patient's normalized gene expression across all genes) and higher disease severity (defined in the primary study as an aggregated measure of percutaneous O_2 saturation, respiratory rate, subcostal retractions, general appearance, and auscultation) compared to the patients in the periphery. The same statistical test was used to compare the weighted degree centrality of the genes in each gene cluster.

Results

Patient Clusters. As shown in Figure 1A, the patients had a complex but understandable topology consisting of a majority of the cases on the top, and a majority of the controls at the bottom of the network. In addition, the cases on the top had a core-periphery topology, where there were patients with high overall gene expression in the center (henceforth referred to as the *core cases*), and many patients with low overall gene expression in the periphery (henceforth referred to as the *periphery cases*). Finally, there were four cases (pink diamonds and triangles) that were clustered with the controls at the bottom of the network (henceforth referred to as the *control-like cases*).

The boundaries of the above clusters were quantitatively verified through agglomerative hierarchical clustering. The vertical dendrogram in Figure 1B shows that there were three main subject clusters: the first consisting of 25 *core cases* (10 flu, 15 RSV), the second consisting of 50 *periphery cases* (16 flu, 34 RSV) with 1 control, and the third consisting of 4 *control-like cases* (2 flu, 2 RSV) clustered with 21 controls. These cluster boundaries were used to color nodes in the network to denote cluster membership as shown in Figure 1A.

To test whether the above clusters could have occurred by chance, we measured their clusteredness with respect to random permutations of the data. The subject clustering in the flu/RSV data was significant when compared to 1000 random networks based on variance of the dissimilarities (*flu/RSV* =2.75, Random-Mean=0.88, p<.001 two-tailed test), skewness of the distribution of dissimilarities (*flu/RSV*=5.55, Random-Mean=3.94, p<.001 two-tailed test), and kurtosis of the distribution of dissimilarities (*flu/RSV*=38.69, Random-Mean=25.03, p<.001 two-tailed test).

Furthermore, the weighted degree centrality (sum of edge weights) of the *core cases* (Median=4.55) was significantly different (*U*=49.00, p<.001, two-tailed test) compared to the *periphery cases* (Median=2.52) suggesting that the overall gene expression of the patients in the core was higher compared to those in the periphery. Finally, the disease severity of *core cases* (Median=7) was significantly higher (*U*=261.50, p<.001, two-tailed test) compared to *periphery cases* (Median=2). Finally, there was no significant difference ($\chi^2(2, N=79)=0.86$, p=0.652) in the proportion of flu vs. RSV patients across the three case clusters, suggesting that the gene-based clustering was common across both types of infection.

Gene Clusters. As shown in Figure 1, the genes fell into two clusters, whose boundaries were quantitatively verified through hierarchical clustering. As shown by the horizontal dendrogram in Figure 1B, there was a large cluster of 14 genes (LDLR, HIST2H2AA3, HIST1H1C, HIST2H2AA4, FCGR1A, TRIB1, SIGLEC1, FCGR1B, IFI27, DEFA1, MMP8, GMPR, RNASE2, H1F0) at the top of the network, and a smaller cluster of the 4 genes (FLJ13197, PTGDR, KLRB1, FCER1A) at the bottom. Based on the results previously published on the same data, the cluster of 14 genes consisted of all up-regulated genes, whereas the cluster of 4 genes consisted of all down-regulated genes. The bipartite network also revealed the inter-cluster relationships: the median gene expression of the 14 genes of the 25 *core cases* (Median=4.22) was significantly higher (*U* =16, p< .001, two-tailed test) compared to the 50 *periphery cases* (Median=1.95). This pattern can also be seen in the high expression values (shown in mostly red cells) in the upper left-hand corner of the heatmap in Figure 1B.

The clusteredness of the above gene clusters was significant when compared to 1000 random networks based on variance of the dissimilarities (flu/RSV=2.91, Random-Mean=0.24, p<.001 two-tailed test), skewness of the distribution of dissimilarities (flu/RSV=2.01, Random-Mean=0.80, p<.001 two-tailed test), and kurtosis of the distribution of dissimilarities (flu/RSV=7.81, Random-Mean=3.16, p<.001 two-tailed test).

Discussion

The bipartite visualization and quantitative verifications revealed not only a strong separation of the cases from the controls, but also a core-periphery topology for the cases. This complex but understandable topology helped to identify three possible subphenotypes and their potential pathways. (1) The *core cases* have significantly higher expression of 14 up-regulated genes, which included 4 histone genes, 4 genes with to date have unknown function in antiviral response, and 6 immune-related genes each of which has a well-known non-overlapping antiviral function. The latter set included **RNASE2** which induces direct damage to viruses, **IFI27** and **DEFA1** which produce specific and general microbicidal protein responses respectively, **FCGR1** which among a multitude of immune functions is primarily involved in activation of the monocyte, macrophages and dendritic cells for efficient antigen presentation, **SIGLEC1** a type I interferon dependent sialoadhesin, and **MMP8** a tissue remodeling enzyme (Collagenase 2). An Ingenuity Pathway Analysis (IPA) of the 14 genes suggested an indirect but strong interferon signature including TNFα and IL-6 cytokines involved in antiviral and innate inflammatory responses. Because the *core cases* also had a significantly higher disease severity score, we hypothesize that these patients represent a distinct at-risk subphenotype that are hyper responsive to pathways targeted to viral clearance, and possibly carry a risk for long-term epithelial cell damage. (2) The *periphery cases* have a medium expression of all 18 genes and therefore suggest a second subphenotype with a subdued anti-viral response relative to the above hyperresponders. (3) The *control-like cases* have a high expression of 4 down-regulated genes, and low expression of the 14 up-regulated genes, and therefore mirror the expression patterns in uninfected controls. The results therefore suggest that the down-regulation of these 4 genes indicates a "protective" phenotype making them similar to the uninfected controls. Existing literature on these genes provide some confirmatory evidence. While the exact role of the high-affinity receptor which binds to the constant portion of IgE (**FcER1**) is unknown in viral pathogenesis, SNPs included on this gene have been shown to be associated with severe RSV disease[5]. Additionally, **KLRB1** which has been shown

to have inhibitory functions on natural killer (NK) cells[16] was downregulated, suggesting an enhanced antiviral response in patients resembling the immune response of controls. Finally, **PTGDR** a receptor important in mast cell function was downregulated, but the exact role of this receptor in viral infection is still unknown. Overall, *control-like cases* suggests a third subphenotype which have a "just enough" response to the virus, without overt stimulation of virally induced genes, and therefore potentially with reduced bystander damage.

One could argue that the above result could also be the result of the progression of infection over time. For example, the *core cases* could be at the peak of infection, the *periphery cases* could be later in the infection, and the *control-like cases* could be recovering from the infection. However, an additional analysis revealed that the 3 case clusters were not significantly different ($H(2, N=79)=2.56$, $p=0.278$) in time of sample collection after hospitalization. There is of course the possibility that the children were infected at very different times before hospitalization, but controlling such a variable is practically impossible in the analysis of human infections. Therefore, we provide two explanations for why sample collection time is probably not an adequate explanation for the results: (1) Because all case samples were collected from patients that were hospitalized indicating severe illness, a resolution of such severity in the short time window of 42-72 hours is unlikely to occur. (2) The gene expression changes in the PBMCs of the patients suggest a specific induced innate immune response (e.g., Toll-like receptor) to viruses. Such signaling pathways (which induce interferon secretion and contribute to anti-viral immunity) last several days which exceeds the sample collection time window in this study. We therefore propose that the three case clusters are more likely the result of inherent host differences in anti-viral responses, and therefore represent distinct subphenotypes.

Conclusions and Future Research

Several epidemiological, clinical, and genetic risk factors have been examined to identify children at risk for severe acute RSV and influenza infection, and for long-term sequelae. However, to the best of our knowledge, no study has applied multivariate methods to analyze gene expression data from naturally-infected children with flu or RSV with the goal of identifying the heterogeneity of their host response and the respective pathways involved. Here we presented a multivariate analysis of gene expression in human data using bipartite networks.

We believe our study makes three biological and methodological contributions. **(1)** We have shown evidence for the existence of three subphenotypes that are common to flu and RSV. While there might be other subphenotypes and underlying pathways that are unique to each disease, we were specifically interested in subphenotypes and pathways that were common to both diseases, with the goal of providing insights into future therapeutic targets that address multiple types of respiratory infections. The study therefore has helped to identify data-driven hypotheses for subphenotypes that can be tested in future studies. **(2)** We have provided biological inferences for the genesis of the three subphenotypes, and argued why time of data collection alone is not an adequate explanation for the results. **(3)** We have demonstrated the utility of bipartite networks to reveal a complex but understandable combination topology consisting of distinct clustering, in addition to a core-periphery topology. Such an understanding of relationships in data is difficult using unipartite methods such as *k*-means (which can identify for example either patient or gene clusters but not both simultaneously), or even bipartite heatmaps with dendrograms (shown in Figure 2B), which are more useful for confirming a topology, rather than for discovering complex associations[8]. Furthermore, methods like modularity[11] (used to identify disjoint clusters in networks) are also not designed to discover such combination topologies, therefore demonstrating the advantages of bipartite network visualizations to guide in the comprehension of complex multivariate associations, and in the selection of appropriate quantitative methods for verification. Given the importance of visualizations to detect such complex topologies, our current research is examining visualizations for "big data" containing hundreds of thousands of patients and variables.

A limitation of our study is that we examined only those genes that were common to both infections, and there might be subphenotypes which are unique to each infection type. Furthermore, we analyzed subphenotypes in only one dataset. Therefore our future research aims to examine genes that are specific to each infection type from the same dataset, and test them in a new dataset in collaboration with the authors (who collaborated in the current study) of the primary study. In that respect, this project demonstrates the promise of the *Open Science*[17] movement, where publicly available data not only enables new hypotheses to be generated from existing data, but can also motivate new interdisciplinary collaborations among researchers who would not have ordinarily been motivated to work together. Such collaborations should help accelerate discoveries in complex phenomena such as disease subphenotypes and their pathways, with the goal of translating them into effective and well-targeted therapeutics.

Acknowledgements

This research was supported in part by a grant from IHII, UTMB, and NIH UL1TR000071 UTMB CTSA (ARB).

References

1. World Health Organization. Acute Respiratory Infections. http://www.who.int/vaccine_research/documents/ARI07062010_2.pdf. Accessed August 28, 2013.
2. Sigurs N, Bjarnason R, Sigurbergsson F, Kjellman B. Respiratory syncytial virus bronchiolitis in infancy is an important risk factor for asthma and allergy at age 7. Am J Respir Crit Care Med. 2000 May;161(5):1501-7.
3. Welliver RC. Review of epidemiology and clinical risk factors for severe respiratory syncytial virus (RSV) infection. J Pediatr. 2003 Nov;143:S112-7.
4. Bhat N, Wright JG, Broder KR. Influenza-associated deaths among children in the United States, 2003-2004. N Engl J Med. 2005 Dec 15;353(24):2559-67.
5. Janssen R, Bont L, Siezen CL, Hodemaekers HM, Ermers MJ, Doornbos G, van 't Slot R, Wijmenga C, Goeman JJ, Kimpen JL, van Houwelingen HC, Kimman TG, Hoebee B. Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. J Infect Dis. 2007 Sep 15;196(6):826-34.
6. Vareille M, Kieninger E, Edwards MR, Regamey N. The airway epithelium: soldier in the fight against respiratory viruses. Clin Microbiol Rev. 2011 Jan;24(1):210-29.
7. Ioannidis I, McNally B, Willette M, Peeples ME, Chaussabel D, Durbin JE, Ramilo O, Mejias A, Flaño E. Plasticity and virus specificity of the airway epithelial cell immune response during respiratory virus infection. J Virol. 2012 May;86(10):5422-36
8. Bhavnani SK, Bellala G, Victor S, Bassler KE, Visweswaran S. The Role of Complementary Bipartite Visual Analytical Representations in the Analysis of SNPs: A Case Study in Ancestral Informative Markers. JAMIA (2012) 19:e5-e12.
9. Bhavnani SK, Victor S, Calhoun WJ, Busse WW, Bleecker E, Castro M, Ju H, Brasier AR. How Cytokines Co-occur across Asthma Patients: From Bipartite Network Analysis to a Molecular-Based Classification. Journal of Biomedical Informatics, 44 (2011) S24–S30.
10. Bhavnani SK, Drake J, Bellala G, Dang B, Visweswaran S, Olano JP. How Cytokines Co-occur across Rickettsioses Patients: From Bipartite Visual Analytics to Mechanistic Inferences of a Cytokine Storm. AMIA Summit on Translational Bioinformatics, 2013.
11. Newman M. Networks: An Introduction. Oxford University Press; 2010.
12. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci U S A. 2007 May 22;104(21):8685-90.
13. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008 Apr;18(4):644-52.
14. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Information Processing Letters. 1989;31(1):7-15.
15. Nooy W, Mrvar A, Batagelj V. Exploratory Social Network Analysis with Pajek. New York, NY: Cambridge University Press, 2005.
16. Pozo D, Valés-Gómez M, Mavaddat N, Williamson SC, Chisholm SE, Reyburn H. CD161 (human NKR-P1A) signaling in NK cells involves the activation of acid sphingomyelinase. J Immunol. 2006 Feb 15;176(4):2397-406.
17. Molloy JC. The Open Knowledge Foundation: Open Data Means Better Science. 2011 PLoS Biology 9.

Use of RxNorm and NDF-RT to normalize and characterize participant-reported medications in an i2b2-based research repository

Colette Blach[1], Guilherme Del Fiol[2], MD, PhD, Chandel Dundee, RN[1], Julie Frund[1], Rachel Richesson, PhD[1], Michelle Smerek[1], Anita Walden[1], Jessica D. Tenenbaum, PhD[1]

[1]Duke University, Durham, NC, [2]Univerity of Utah, Salt Lake City, UT

Abstract

The MURDOCK Study is longitudinal, large-scale epidemiological study for which participants' medication use is collected as free text. In order to maximize utility of drug data, while minimizing cost due to manual expert intervention, we have developed a generalizable approach to automatically coding medication data using RxNorm and NDF-RT and their associated application program interfaces (APIs). Of 130,273 entries, we were able to accurately map 122,523 (94%) to RxNorm concepts, and 106,135 (85%) of those drug concepts to nodes under the Drug by VA Class branch of NDF-RT. This approach has enabled use of drug data in combination with other complementary information for cohort identification within an i2b2-based participant registry. The method may be generalized to other projects requiring coding of medication data from free-text.

Introduction and Background

Standardized drug terminologies are useful to facilitate data sharing, and to ensure semantic interoperability across organizations [1]. Even when medication data is not collected in a coded manner, RxNorm[1] and the VA National Drug File Reference Terminology (NDF-RT)[2] have, with certain caveats [1], proven useful for normalizing both structured and free text data from electronic health records (EHRs) [2, 3]. We have taken a similar approach to mapping participant-provided drug information to RxNorm and NDF-RT to enable cohort identification using the i2b2 platform [4].

Medication information is an important facet of a person's medical history. Medication data from EHRs may be limited to prescriptions taken as an inpatient or prescribed by a clinician at the health care facility in question. In addition, over-the-counter medications, vitamins, and supplements may not be included. The work described here was done in the context of the MURDOCK Study, a long-term epidemiological study aimed at reclassifying human health and disease based on molecular mechanism rather than the macroscopic observations that have been used for hundreds of years [5]. Participants in the study provide blood and urine biospecimens, along with self-reported clinical, medication, demographic, lifestyle, and medical history data. They also provide consent to annual follow up and access to their EHRs. The MURDOCK Study has an advantage over EHR-only projects in that medication data is to be provided both from EHRs and as self-reported information. To date, approximately 9600 participants have been enrolled out of an ultimate goal of 50,000 participants.

Here we describe the successes, limitations, and caveats of coding free text medication data using RxNorm and NDF-RT in the context of patient-reported information and contrast these factors with those described previously using EHR-derived medication data [3].

Methods

A graphical overview of the method described in this manuscript is given in **Error! Reference source not found.**. Participant-reported medications were collected by study staff as free text (A) annually for all participants in a community-based registry [6]. These free text medication entries were mapped, where possible, to RxNorm CUIs (Concept Unique Identifiers) using the National Library of Medicine (NLM)'s RxNorm REST API (B). A hierarchical structure of drug classes was developed based on the NDF-RT's "Drug Products by VA Class" subtree (C) by replacing drug concepts containing attributes such as route and dosage (e.g., ibuprofen 200mg oral) with ingredient concepts (e.g., ibuprofen) and brand name concepts (e.g., Advil) (D). Multiple or single ingredient RxNorm concepts in our dataset were mapped to this hierarchy, followed by brand name terms as leaf nodes to the ingredient or sets of ingredients.

---

[1] http://www.nlm.nih.gov/research/umls/rxnorm/ (Accessed 10/9/13)

[2] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/ (Accessed 10/9/13)

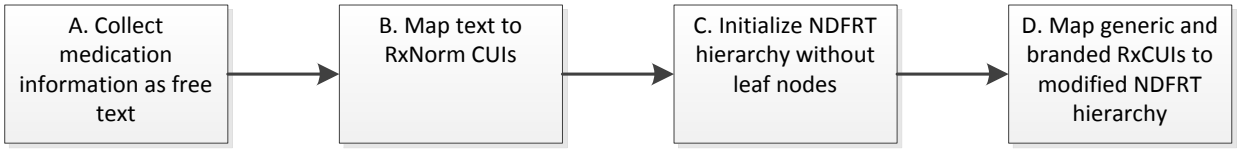


Figure 1: Methods overview.

Mapping free text medication data to RxNorm

Participants were instructed to list medications by generic or brand names, leaving out other attributes such as route, strength, and form. To facilitate the accurate collection of medication information, participants were requested to bring their medications to the enrollment visit. The NLM RxNorm API was run on free text data from all participants enrolled in the MURDOCK registry as of 7 June 2013. A total of 130,273 medication entries were present, representing 18,924 unique terms reported by 9432 participants. An additional 2,579 entries (16 unique) included non-medication terms, e.g. "no medications", "can't afford drugs", etc. and were excluded from the analysis.

An attempt was made to detect perfect string matches first (http://rxnav.nlm.nih.gov/REST/drugs?name=*value*). For those that did not return a perfect match, the approximate match resource was used, e.g. http://rxnav.nlm.nih.gov/REST/approx?term=aspirin.[3] This resource returns 0 or more RxNorm terms that match the input string, in this case aspirin. The output includes a set of potential matches along with their RxCUI, score (an integer between 1 and 100 that measures the similarity between the input string and the candidate RxNorm term), and rank. Details on the approximate match algorithm are available elsewhere [7]. In the case of multiple matches, a winner was chosen using the following rules:

- Highest score, or in case of tie:
- Non-proprietary > proprietary source terminology (RxNorm content is derived from 11 source terminologies, some of which are proprietary.)
    - Among non-proprietary RxNorm concepts > NDF-RT concepts
        - RxNorm concept type: ingredient name (IN) > brand name (BN)

Results were categorized as follows:

- Category A: Perfect match (no score assigned)
- Category B: Score == 100 for exactly 1 term, and that one is non-proprietary
- Category C: Score == 100 for more than 1, and winner is non-proprietary
- Category D: Score == 100 for proprietary only (whether 1 or more)
- Category E: Match score < 100
- Category F: No match found

Category E was further divided into E3: scores (s) < 50; E2: $50 \leq s < 75$; and E1: $s \geq 75$.

Category A matches were inspected visually and determined to be 100% accurate. For each of the remaining categories and sub-categories (B, C, D, E1-E3), 100 term mappings were selected at random and reviewed by an analyst with clinical expertise who determined whether the mapping was correct. In addition, all Category B matches (n=17) were reviewed. Therefore, 517 approximate matches were manually reviewed and evaluated for accuracy.

Mapping to an NDF-RT-based hierarchy

In order to be able to query for drugs by class, as opposed to name or ingredients, it was necessary to map RxNorm terms to another terminology that included categorization. NDF-RT's "Drug Products by VA Class" hierarchy has

[3] The approxMatch function used here was released in September 2011. A similar function, approximateTerm, was released in May 2013 that gives similar results but provides additional output control, e.g. maximum number of entries returned.

been used for this purpose. NDF-RT leaf nodes, however, include dose and route information, e.g. “ibuprofen, 20mg tablet, oral”. In contrast, the medication data collected (and thus the corresponding RxCUIs) generally included only drug names, not strength or route, e.g. “ibuprofen, 200mg tablet, oral.” It was therefore necessary to rebuild the NDF-RT hierarchy down to the drug name level, without route or dose information. This revised version was designed to include brand names relationships in the hierarchy so that, for example, a query for participants taking ibuprofen would also return participants who reported taking Advil, Midol, Motrin, etc. This was accomplished through the following steps:

1. For each of the original leaf nodes (e.g., carbamazepine 100mg tab, chewable and carbamazepine 100mg/5ml susp)
    a. Identify the term type (TTY) of each RxCUI for each NDF-RT leaf node.
    b. “Walk” the NDF-RT ontology back to ingredient[s]. This is done by taking advantage of the relationships between different term types within RxNorm, e.g. clinical drug form (SCDF) has a “has_ingredient” relationship to ingredient (IN).
        i. If the drug comprised multiple ingredients, append a multiple-ingredient child node onto the initial hierarchy.
        ii. For those mapped only to a single ingredient, walk to ingredient and create a child node under NDF-RT.
    c. If the RxNorm concept had no relationships:
        i. Map NDF-RT product component[s] (converting from precise ingredient (PIN) to ingredient (IN) if applicable) to [multiple] ingredients and create a node if one does not already exist.
        ii. Parse drug name and use string match to identify ingredient or brand name and create a node if one does not already exist.
2. Create child nodes of the nodes created above for each brand associated with that ingredient (or set of ingredients).

In the modified NDF-RT VA Drug Class hierarchy, RxCUIs for MURDOCK drug entries corresponded directly to the Ingredient (or Multiple Ingredient) and Brand Name nodes.

Results

Mapping free text to RxNorm

Out of 9432 participants, 8356 indicated taking one or more medications (including OTC medications, vitamins, and supplements) at one or more time points. The terms entered largely did not include dosages or delivery mechanism. This resulted in 130,273 total (18,924 unique) drug name entries. As illustrated in Figure 2, 99,538 entries (76%) of terms were perfect matches. On the other hand, the majority of the unique terms fell into category E (14,114 out of 18,924; 75%). This was to be expected as there are a number of different incorrect ways to spell a given drug name, and only one correct way.

Based on manual expert review of a random sample from each category, accuracy rates were determined to be 100% for categories A-D, and 98%, 92%, and 62% for E1, E2, and E3 respectively. Thus 104,269 (80%) of the total terms could be mapped to concepts in RxNorm with near 100% accuracy. Another 5332 (4%) could be mapped with approximately 98% accuracy, and 14,162 (11%) with accuracy of approximately 92%.

Choosing the appropriate threshold level below which to reject matches was subjective. While choosing a score of 100 or 75 would have maintained the highest level of accuracy, a significant number of entries (14,162 term entries; 7060 unique terms) fell into category E2. It was therefore decided to include through category E2 (scores ≥ 50) to maximize coverage, and keep the score to indicate confidence level.

Using a cutoff threshold of ≥ 50 (i.e. excluding only category E3), we were able to map 123,763 terms (**Error! Reference source not found.**). Extrapolating from our established rates of accuracy in each category, ~94% (122,523) of the total terms were mapped correctly, with 5% unmapped, and a "false positive rate" (i.e. mapped, but incorrectly) of approximately 1% (1,240 entries). Manual review suggests that missing and incorrect mappings were primarily over-the-counter medications, vitamins, minerals, and other supplements.

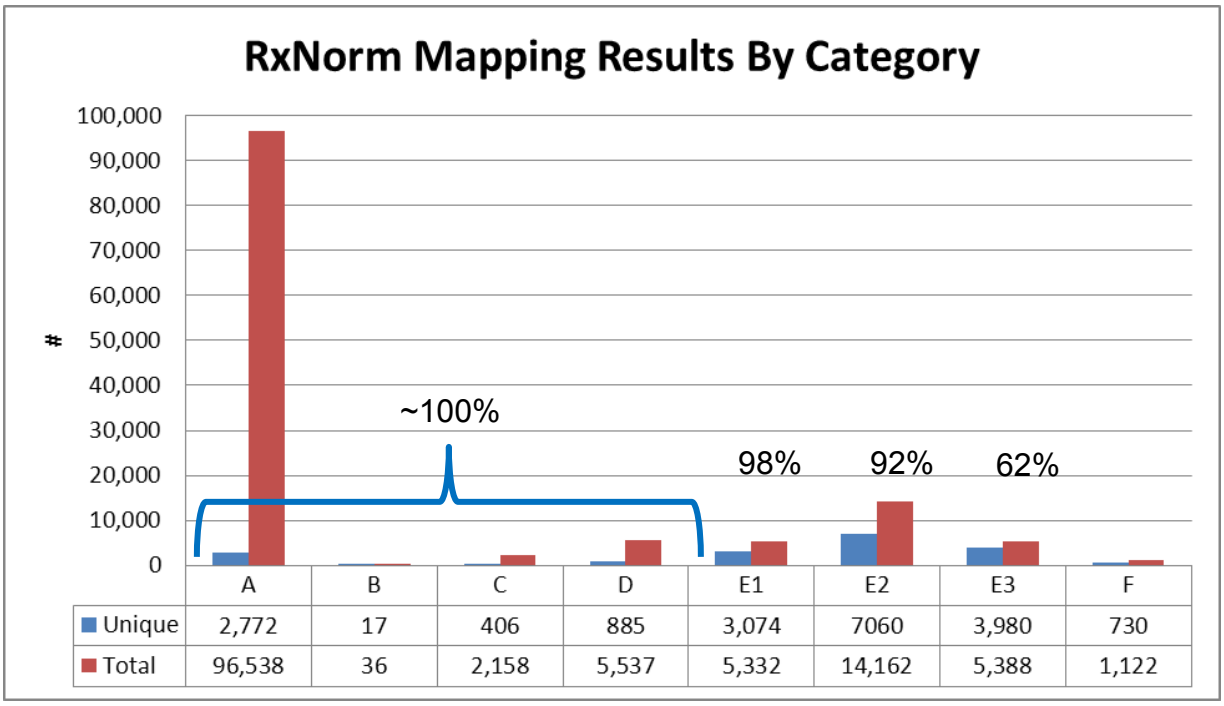


Figure 2: Distribution of input terms across scoring categories. A. Perfect matches; B: Score == 100 for exactly 1 term, and that one is non-proprietary; C: Score == 100 for more than 1, and winner is non-proprietary; D: Score == 100 for proprietary only (whether 1 or more); E1: 75 ≤ Match score < 100; E2: 50 ≤ Match score < 75; E3 Match score < 50; F: No match found.

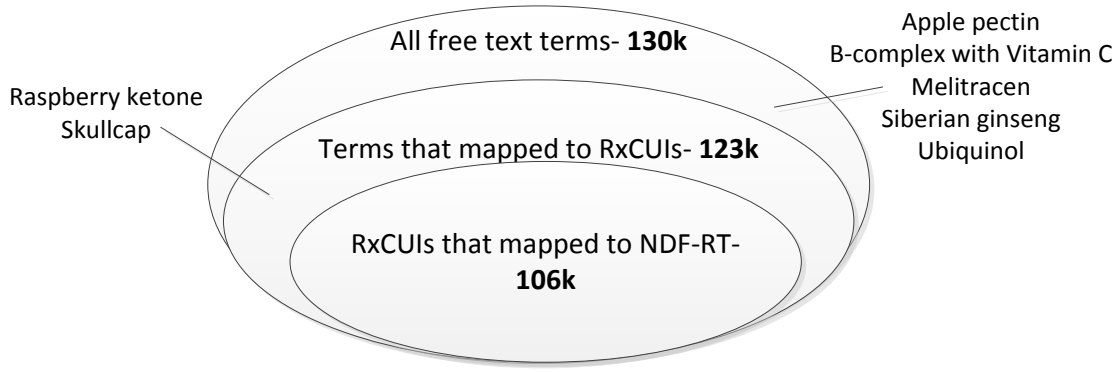


Figure 3: Successful mapping rates from terms to RxCUIs and from RxCUIs to NDF-RT classes. (Not to scale.)

Mapping RxNorm to NDF-RT

104,653 (2158 unique) of the 123,763 (3137 unique) RxNorm terms (85% of the terms; 69% of the unique terms) obtained from mapping text to RxCUIs could be mapped directly to VA Class (**Error! Reference source not found.**). Many of the remaining 19,110 (969 unique) terms are not within scope for the NDF-RT, e.g. medical devices and brand names. Additional transformations, e.g. further tree walking, parsing and remapping, brought these numbers up to 106,135 (86%) of the total terms, representing 2790 (89%) unique RxCUIs.

Discussion and Conclusion

In this study, we designed and evaluated a method for automatically coding free-text medication data into standard terminologies. The method has some important strengths: 1) fully-automated; 2) accurate; 3) leverages publicly available tools and standard terminologies; and 4) can be applied to other uses cases that also involve free-text medication data.

In the planning phase of this project, one proposed option for drug coding was to use a hybrid model of automated and manual coding. However, cost estimates for professional coding at the level that is commonly performed for clinical trials were on the order of $1 million for full coding of 50,000 participants, including 25% manual coding. In evaluating coverage, accuracy, and cost of the mapping approach, it is important to consider the use case for which this system is intended. In the case of the MURDOCK Study, ascertainment of concomitant medications is a means to an end, namely identifying specific participant cohorts for follow-up studies. Importantly, this data itself is not to be used for, e.g. monitoring severe adverse events or FDA submission. Therefore, the automated coding approach we have described is a viable solution to standardization of free text drug data.

To evaluate the performance of the RxNorm API matching algorithms, Peters et al. mapped clinical drug terms (e.g., metoprolol succinate 200mg tab) from different drug formularies to RxNorm concepts [7]. The evaluation resulted in a recall of 61% to 76% of unique terms compared to 75% in our study, despite the fact that our data set had a large number of misspelled terms. This similar performance demonstrates the robustness of the RxNorm matching algorithm for different data sources, further indicating the generalizability of our method to other use cases and data sets.

Previous efforts to code existing medication data have been performed primarily on data collected through clinical care, whether structured or free text [3]. Our efforts to apply the RxNorm to participant-reported data adds additional complications in that fewer details were included (i.e. no dose or route, making mapping to NDF-RT more difficult) and participants are less familiar with drug names and spellings. On the other hand, the designation of medication data is more straightforward in this context than previously reported use of NLP to extract medication information from notes in EHRs. Our mapping success rate therefore shows immediate promise for easy application of RxNorm and its related APIs to participant-reported medications in research contexts.

Limitations

By relying on NDF-RT, we are constrained to using only the class assigned to each drug in the legacy VA Drug class system. However, many drugs fall into different categories and/or are used for multiple different indications. In order to use drug information to identify cohorts, particularly without knowledge of dosage, it was critical to also collect the reason for taking the drug. For example, the “antiparkinson agents” class of drugs in NDF-RT, is a reasonable start for identifying a cohort of Parkinson’s patients. However, because antiparkinson agents may be used for other conditions, e.g. restless leg syndrome, it is necessary to search both by drug type and reason taken in order to avoid false positives.

The accuracy threshold chosen in this study might not apply to all other systems or contexts. For example, investigators interested in the use of vitamins, minerals, and dietary supplements might have lowered the acceptable threshold, as those types of products tended to have lower matching scores. In contrast, a higher threshold might be justified for a use case in which false positives were particularly problematic.

In addition to the evaluation above, which focuses on the proportion of total entries and unique entries that were accurately mapped to the drug terminologies, two additional questions one might ask are what percentage of people who tried to report taking a given drug do we know reported taking it (e.g. if “asprin” mapped correctly to aspirin, but “asparin” did not, we would only know about those who misspelled it one way and not the other), and what percentage of people do we think reported taking a drug who in actuality did not (e.g. askarin was mapped to aspirin, but the participant actually misspelled the brand name Akarin)? However, these questions are highly drug specific. This type of analysis may be done in the future if it becomes relevant for a specific use case.

Future work

The Anatomical Therapeutic Chemical (ATC) is a terminology maintained by the WHO Collaborating Centre for Drug Statistics Methodology. It offers categorization analogous to that provided by NDF-RT. During the course of the work described here, the Anatomical Therapeutic Chemical (ATC) classification system was mapped to RxNorm. Future work will include evaluation of ATC as an alternative to NDF-RT, as it appears to address some shortcomings of NDF-RT, for example providing more than one class for a given drug (NDF-RT currently does this only for a very few cases), or the presence of certain drug classes not present in NDF-RT (e.g., selective serotonin reuptake inhibitors). It has been shown that the overlap in drug-class pairs between NDF-RT and ATC is poor [8]. Time and user input will determine whether NDF-RT is sufficient for the task at hand in the MURDOCK registry, or whether an alternative terminology is needed.

Acknowledgements

This work was funded by Duke's Clinical and Translational Science Award UL1RR024128 and a gift from David H. Murdock.

References

1. Pathak, J. and C.G. Chute, *Further revamping VA's NDF-RT drug terminology for clinical research.* J Am Med Inform Assoc, 2011. **18**(3): p. 347-8.
2. Palchuk, M.B., et al., *Enabling Hierarchical View of RxNorm with NDF-RT Drug Classes.* AMIA Annu Symp Proc, 2010. **2010**: p. 577-81.
3. Pathak, J., et al., *Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project.* AMIA Annu Symp Proc, 2011. **2011**: p. 1089-98.
4. Murphy, S.N., et al., *Integration of clinical and genetic data in the i2b2 architecture.* AMIA Annu Symp Proc, 2006: p. 1040.
5. Tenenbaum, J.D., et al., *The MURDOCK Study: a long-term initiative for disease reclassification through advanced biomarker discovery and integration with electronic health records.* Am J Transl Res, 2012. **4**(3): p. 291-301.
6. Bhattacharya, S., et al., *The Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study Community Registry and Biorepository.* Am J Transl Res, 2012. **4**(4): p. 458-70.
7. Peters, L., J.E. Kapusnik-Uner, and O. Bodenreider, *Methods for managing variation in clinical drug names.* AMIA Annu Symp Proc, 2010. **2010**: p. 637-41.
8. Mougin, F., A. Burgun, and O. Bodenreider. *Comparing drug-class membership in ATC and NDF-RT*. in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. 2012: ACM.

Capturing cancer initiating events in OncoCL, a cancer cell ontology

Mary E. Dolan
Department of Bioinformatics and Computational Biology
The Jackson Laboratory, Bar Harbor, ME 04609

Abstract

We have developed an ontology, OncoCL, to classify cancer cells and provide a framework for consistent annotation of cancer-associated data from conventional surgical pathology and cancer molecular biology for the purpose of access, comparison, and analysis. The cell type ontology, CL, describes normal cell types and was not designed to capture the pathology of cancer cells. OncoCL builds upon CL, as a canonical cell (represented in CL) undergoes oncogenic change and tumorigenesis with the acquisition of the cancer hallmarks described by Hanahan and Weinberg.

The characterization of cancer initiating cells and cancer initiation events present particular challenges – for example, the representation of the self-renewal and differentiation potential of cancer stem cells compared with those of canonical (normal) stem cells. But we know that the distinction of high-risk precursor lesions with a high likelihood of developing into cancer, compared with indolent disease, depends on the synthesis of complex, heterogeneous data related to cancer initiating cells. OncoCL is a flexible resource specifically developed to integrate these diverse data through the reuse of a number of other biomedical ontologies. This work will present the problems we encountered capturing cancer initiating events and the solutions we implemented to address them.

Automatic Gene Prioritization in Support of the Inflammatory Contribution to Alzheimer's Disease

Stephanie K Furniss, MS[1], Robert Yao[1], Graciela Gonzalez, PhD[1]
[1]Arizona State University, Phoenix, AZ

Abstract
This research seeks to extend the process of novel therapeutic gene target discovery for the treatment of Alzheimer's disease (AD). Gene-gene and gene-pathway annotation tools as well as human analysis are used to explore likely connections between potential gene targets and biochemical mechanisms of AD and associated genes. Rule-based annotation systems, such as GeneRanker, can be applied to the continuously growing volume of literature to extract relevant gene lists. The subsequent challenge is to abstract biological significance from associated genes to aid in discovery of novel therapeutic gene targets. Automatic annotation of genes deemed significant by data-driven assays and knowledge-driven analysis is limited. Therefore, human analysis is still crucial to exploring novel gene targets and new disease models. This research illustrates a method of analysis of an extracted gene list which lead to the discovery of KNG1 as a possible therapeutic target, suggests a connection between inflammation and AD pathogenesis.

Introduction

Alzheimer's disease (AD) is a neurodegenerative disease that causes progressive decline in cognitive functions and impairment of memory. In 2013, AD affects an estimated 5.2 million Americans, and costs of care to the country are projected to reach $203 billion (1). There is no method to prevent, suspend or reverse disease progression; as a result prevalence of AD in the United States is expected to continue its increasing trend. Given its impact on public health, current efforts in AD research are focused on finding as many of the genetic and molecular bases of the disease for the purpose of eventual treatment. Genome-wide association studies (GWAS) have identified novel genetic variants associated with AD (2, 3). But conventional GWAS approaches may not be able to detect some genetic variation or gene-gene interactions (4). This study in particular seeks to extend the process of novel therapeutic gene target discovery.

Published literature is a rich source of genomic information, but the volume of published literature is too large for a biomedical researcher, or group of researchers, to remain up-to-date. Where biomedical researchers may use molecular assays to identify a set of disease-associated genes or genes of interest, biomedical text-mining researchers use integrative gene annotation methods to produce an expanded network of genes associated with a disease. The amount of genes deemed significant from automatic data-driven (large-throughput) assays and knowledge-driven analysis is currently limited. Thus, human analysis is still crucial to exploring novel gene targets and new disease models.

In order to accomplish this, previously manually curated GWAS data related to AD are used as a seed set for the automatic detection of novel and potentially relevant gene targets contained within the literature. Using the results of these experiments, automatic analysis methods more robust than basic statistical methods were employed to extract meaningful genomic relationships. Such work can aid in discovery of novel therapeutic gene targets by informing bench researchers of AD-relevant genes that may be worthy of further investigation or provide evidence to support current hypotheses. It can also be used by biomedical text-mining researchers to guide development of an automated tool that incorporates gene-pathway associations to facilitate understanding of a large gene list, filtering for relevance to research interests, and discovery of novel gene-disease associations via pathway analysis.

Methods

Gene list from AlzGene.org database, which is an actively maintained field synopsis of genetic association studies in Alzheimer's disease, defined the gold standard curated gene list—AlzGene623. AlzGene623 is composed of 623 genes collected from the AlzGene.org database (5). The second seed list named GwasList consists of 547 unique genes from a list of significant SNPs from a GWAS study completed by a collaborating biomedical researcher. The third seed list, SearchByAD, consists of 742 genes and is a product of GeneRanker's gene-disease annotation function.

GeneRanker is an automated gene prioritization tool developed by the Diego lab at Arizona State University to aid biomedical researchers in the discovery of novel gene therapeutic targets. GeneRanker's integrative method examined gene-gene annotations to expand on each seed gene list, resulting in an enriched gene list. It has been evaluated in the context of brain cancer research and atherosclerosis (6, 7). GeneRanker's

computational method combines data extracted from the literature and from curated sources, such as Genetic Association Database and NCBI Gene database.

The four uniquely-developed seed lists were entered into GeneRanker. Each provided a subsequent enriched gene list after gene-gene annotations were examined through an integrative method. Table 1 provides a tabular comparison of the seed lists to show the number of overlapping genes between lists. Combined, the seed lists comprise 1523 unique genes (SeedGenes).

	GwasList	SearchByAD	AlzGene623
GwasList	547	27	25
SearchByAD		742	353
AlzGene623			623

Table 1. Comparison of the three seed lists. Table gives the number of overlapping (shared) genes between two seed lists.

Each enriched gene list was plotted on a computational score versus computational rank graph. A cut off point located where the curve began to flatten was marked (AlzGene623 enriched gene list and cut-off point is graphed in Figure 1). Genes right of the cut off threshold were disregarded due to their low rank and score. Genes to the left of the cut-off point (highly-ranked) may contain potential novel target genes as well as seed genes. For example, of the 990 highly-ranked genes from AlzGene623 expanded gene list 23.2% are in AlzGene623, providing 760 genes for further consideration as novel therapeutic targets. This process was followed for seed lists GwasList and SearchByAD as well resulting in, respectively, 609 potentially novel of the 645 highly-ranked and 711 potentially-novel of the 997 highly-ranked genes. 236 genes were common between these three potentially-novel highly-ranked genes. The 236 extracted potential gene targets (ExPot list) will be further examined using gene-pathway analysis.

GATHER was used for KEGG Pathway (pathway) enrichment analysis of gene lists. GATHER is a tool that integrates various forms of available data and applies a statistical model that quantifies the significance of functional associations (8). It was developed to be used by biomedical researchers to understand the function of a group of genes by showing the user the annotations that distinguish the genes entered from other genes in the genome (8). Enrichment analysis are favorable because biological processes are made up of a group of genes, as opposed to an individual single gene (9). The enriched gene lists were entered into GATHER, which then returned associated gene functions and biological pathways annotated by KEGG Pathway.

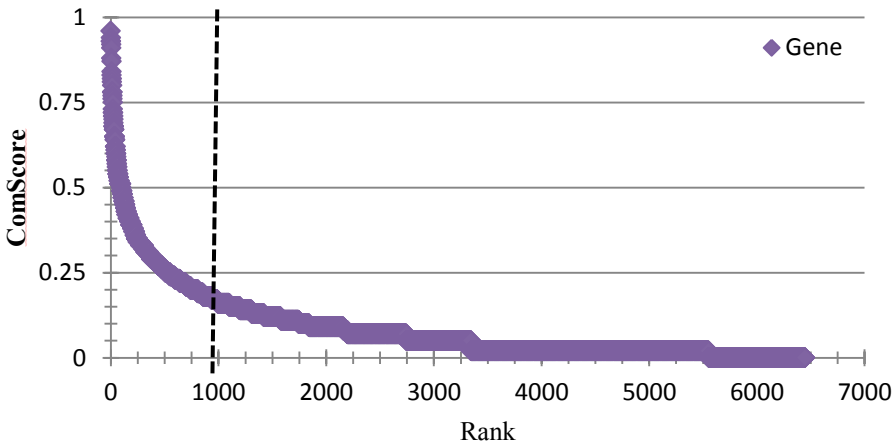


Figure 1. Using GeneRanker, the AlzGene623 seed list expanded to 6449 genes. 990 highly-ranked genes lie to the left of the cut-off point (marked by the dashed line). 23.2% of the highly-ranked genes are in the seed set (AlzGene623), providing 760 genes for further consideration as novel therapeutic targets.

Results

Pathway analysis used significant pathways associated with all seed genes (SeedGenes list), the ExPot list, the combination of the SeedGenes and ExPot (AllGenes list), as well as the four individual seed lists (Table 2). SeedGenes had nine significant ($p<.01$) pathways (Table 2 a-i). ExPot list had 15 significant pathways (Table 2 c, e, g, j-q, s-v). AllGenes had 18 significant pathways (Table 2 a-g, i-s). Table 2 gives the number of genes per associated pathway and the pathway ranking per gene list.

22 of the 1523 SeedGenes were associated with AD pathway (Table 2 a). The ExPot list had zero genes associated to the AD pathway when examined alone and when examined in combination with SeedGenes, in AllGenes (Table 2 a).

Four of the nine SeedGene pathways ranked lower on addition of the ExPot genes (Table 2 a, b, d, f). Three SeedGene pathways increased their ranking on addition of the ExPot genes (Table 2 c, e, i). One did not change ranking (Table 2 g) and one was not ranked in analysis of AllGenes (Table 2 h).

For two pathways there were ExPot genes associated with the pathway when examined in the context of AllGenes, but not when examined in the context of the ExPot list on its own (Table 2 d, f). In one case, 31 genes in AllGenes are associated with the Complement and coagulation cascades pathway. One of the 31 is in the curated AlzGene623 list (CR1). Five of the 31 genes are in the ExPot list (Table 3). Interestingly, these five genes do have a significant association to the complement and coagulation cascades pathway when examined in context of their source gene list, ExPot list.

KEGG Pathway	SeedGenes	AllGenes	ExPot	GwasList	AlzGene623	SearchByDisease
a Alzheimer's disease	22 (1)	22 (3)	--	4 (5)	20 (1)	21 (1)
b Pyrimidine metabolism	2 (2)	2 (5)	--	--	1 (6)	--
c Insulin signaling pathway	46 (3)	74 (2)	28 (3)	--	29 (2)	--
d Neuroactive ligand-receptor interaction	79 (4)	99 (13)	--	19 (4)	--	--
e Apoptosis	33 (5)	49 (4)	16 (10)	--	19 (4)	22 (7)
f Complement and coagulation cascades	26 (6)	31 (16)	--	--	14 (7)	23 (3)
g Calcium signaling pathway	54 (7)	75 (7)	21 (14)	15 (2)	--	--
h Prostaglandin and leukotriene metabolism	16 (8)	--	--	--	--	13 (5)
i Purine metabolism	13 (9)	13 (8)	--	--	3 (3)	1 (2)

Table 2. Subset of the full pathway analysis results from GATHER for seven gene lists examined. The ranking of the pathway is shown in parentheses for each gene list.

Discussion

Three seed lists identified from three unique sources were applied to an automated gene enrichment and prioritization tool to expand on genes already known for the discovery of novel genes related to AD. The overlap of these expanded gene lists were used to identify potential gene therapeutic targets, ExPot list. The ExPot list was further narrowed using pathway analysis of relevant genes. The findings of this study confirm a known association to AD while supporting the potential of the ExPot genes. They also demonstrate an approach to gene- and pathway-disease analysis can lead to a single gene or group of similar genes worthy of further evaluation.

An important, though not surprising, finding with respect to AD is that the pathway analysis of SeedGenes ranked AD as the most significant pathway (Table 2, a) confirming the SeedGenes association to AD. Further, no ExPot gene is associated to the AD pathway and the addition of the 236 ExPot genes dilutes the ranking of the AD pathway (Table 2, a) suggesting confirmation that a novel gene-AD association may lie within the ExPot genes.

For two pathways—complement and coagulation cascade and neuroactive ligand-receptor pathway—there were ExPot genes associated with each pathway when the list was examined in context of AllGenes, but not when

examined in the context of the ExPot list on its own. One of these, the complement and coagulation cascade pathway, had five ExPot genes associated to it (Table 3). This could suggest that the five genes are well-studied and have a weak correlation to the pathway or it could suggest the genes are not well studied, which, if true, increases the likelihood that they too have not been examined in relation to AD. The complement cascade is an indispensable element of the innate immune response (10) and neuroinflammation is believed to be an underlying mechanism in AD, therefore the complement cascade is relevant to AD pathogenesis.

Of the five ExPot genes implicated in the complement and coagulation cascades (Table 3), four are receptors and one, kininogen 1 (KNG1), is a cofactor to coagulation and inflammation. PubMed literature search revealed that a gene-AD association has been examined for some of these ExPot genes (Table 3), consequently removing them from consideration as a potential novel AD therapeutic target. Coagulation factor II (thrombin) receptor (F2R) is not further explored because the lack of evidence connecting AD and coagulation. KNG1 encodes high molecular weight kiniongen protein (HMWK), which plays a role in inflammation, regulation of blood pressure, and coagulation; therefore KNG1 and its precursors are further examined here. As an immune-cell membrane protein, complement component (3d/Epstein Barr virus) receptor 2 (CR2, CD21), is involved in immune responses.

Gene Symbol	Gene Name	Pathway	Examined association with AD?	Citation(s)
BDKRB2	bradykinin receptor B2	Co, Ca, N, Cy	Yes	(11), (12)
C5AR1	complement component 5a receptor 1	Co, N	Yes	(13)
CR2	complement component (3d/Epstein Barr virus) receptor 2	Co	No	
F2R	coagulation factor II (thrombin) receptor	Co, Ca, N	No	
KNG1	kininogen 1	Co	No	
Pathway key: "Co" is Complement and coagulation cascades; "Ca" is Ca^{2+} signaling pathway; "N" is Neuroactive ligand-receptor interaction; "Cy" is Regulation of actin cytoskeleton.				

Table 3. Five ExPot genes associated with the Complement and coagulation cascade pathway.

Cleavage of HMWK results in bradykinin (BK) and cleaved high-molecular-weight kininogen (HKa). HKa, using intracellular signaling pathways, contributes to the pathogenesis of inflammatory diseases by releasing cytokines TNF-α, IL-1β, IL-6, and chemokines IL-8 and MCP-1 from isolated human mononuclear cells (14). HKa may exert antiadhesive effects, thereby regulating leukocyte recruitment into inflamed tissue (15). TNF-α, IL-1β, IL-8, and IL-6 are in the curated gene list therefore providing a strong relationship to genes that have known association to AD.

BK acts through receptors BDKRB2 (ExPot list) and BDKRB1 (16) to mediate activation of proinflammatory signals and regulate cardiovascular processes. A recent study suggested activation of BDKRB1 as a novel therapeutic approach for AD based on evidence that BDKRB1 activation plays an important role in limiting the accumulation of Aβ in AD-like brain possibly through the regulation of activated glial cell accumulation and release of pro-inflammatory mediators (16). No human-model studies of BDKRB1 in association to AD were found in a PubMed search.

As part of the inflammatory response, the complement cascade and KNG1 may deserve greater attention in the search for therapeutic targets for the treatment of AD. In fact, it has been suggested that antibodies to kininogen or peptidomimetics might be a useful and safe therapy in inflammatory diseases or sepsis involving cytokines (17).

Conclusion

Examination of the five ExPot genes associated with the complement and coagulation cascade lead to the novel identification of KNG1 as a potential therapeutic target for treatment of AD. Previous research suggesting that antibodies to kininogen might be a useful and safe therapy in inflammatory diseases involving cytokines (17) further increases the interest and likelihood that KNG1 could be a therapeutic target for treatment of AD.

This research again confirmed GeneRanker, while it also created credibility of the prioritization-extracted gene list through manual review of published scientific literature and automatic annotations.

It is possible that the highly-ranked extracted gene lists include false positives, genes that have been previously studied in association with Alzheimer's but were not in the seed lists or 'noisy' genes. It is important to consider such false positives to examine how our biological knowledge base is driving the gene extraction. For

example, there may be noise from a large variation. As a result, the gene annotation method reaches genes which encode numerous protein kinases and MAP kinases, which are not disease specific.
In this study, pathway analysis was applied to relevant GWAS data (GwasList) in an effort to weed through the noise and narrow down potential gene target lists. Our lab is exploring methods for identifying 'noisy' genes

A gene-enrichment study relies on pathway analysis that is only as good as the functional information providing its pathway definitions. The differences across pathway databases can lead to divergent enrichment analysis results tool to tool. Only one tool, GATHER, was used for gene-enrichment pathways analysis in this study. It may be advisable to incorporate pathway analysis results from another tool, such as DAVID, as well.

Acknowledgements

This research was funded by An Integrative Approach for the Discovery of Potential Therapeutic Targets for Alzheimer's Disease grant awarded to Graciela Gonzalez, PhD, with collaborators Matthew Huentelman, PhD and Eric Reiman, PhD.

References

1. Alzheimer's Association. 2013 Alzheimer's disease facts and figures. Alzheimers Dement. 2013;9(2):208-45.
2. Harold D, Abraham R, Hollingworth P, et al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat Genet. 2009 Oct;41(10):1088-93.
3. Lambert J, Heath S, Even G, et al. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat Genet. 2009 Oct;41(10):1044-9.
4. Morgan K. The three new pathways leading to Alzheimer's disease. Neuropathol Appl Neurobiol. 2011 Jun;37(4):353-7.
5. Bertram L, McQueen M, Mullin K, Blacker D, Tanzi R. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet. 2007 Jan;39(1):17-23.
6. Gonzalez G, Uribe JC, Armstrong B, McDonough W, Berens ME. GeneRanker: An Online System for Predicting Gene-Disease Associations for Translational Research. Summit on Translat Bioinforma. 2008;26:5.
7. Gonzalez G, Uribe JC, Tari L, Brophy C, Baral C, editors. Mining Gene-Disease Relationships from Biomedical Literature: Weighting Protein-Protein Interactions and Connectivity Measures. Pac Symp Biocomput; 2007; Maui, Hawaii.
8. Chang JT, Nevins JR. GATHER: a systems approach to interpreting genomic signatures. Bioinformatics. 2006;22(23):2926-33.
9. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan 1;37(1):1-13.
10. Aiyaz M, Lupton MK, Proitsi P, Powell JF, Lovestone S. Complement activation as a biomarker for Alzheimer's disease. Immunobiology. 2012 Feb;217:204-15.
11. Prediger RDS, Medeiros R, Pandolfo P, et al. Genetic deletion or antagonism of kinin B1 and B2 receptors improves cognitive deficits in a mouse model of Alzheimer's disease. Neuroscience. 2008;151(3):631-43.
12. Mendonsa G, Dobrowolska J, Lin A, Vijairania P, Jong YJ, Baenziger NL. Molecular Profiling Reveals Diversity of Stress Signal Transduction Cascades in Highly Penetrant Alzheimer's Disease Human Skin Fibroblasts. PLoS One. 2009;4(2):e4655.
13. Ager RR, Fonseca MI, Chu SH, et al. Microglial C5aR (CD88) expression correlates with amyloid-beta deposition in murine models of Alzheimer's disease. J Neurochem. 2010;113(2):389-401.
14. Khan MM, Bradford HN, Isordia-Salas I, et al. High-Molecular-Weight Kininogen Fragments Stimulate the Secretion of Cytokines and Chemokines Through uPAR, Mac-1, and gC1qR in Monocytes. Arterioscler Thromb Vasc Biol. 2006 October 1, 2006;26(10):2260-6.

15. Chavakis T, Kanse SM, Pixley RA, et al. Regulation of leukocyte recruitment by polypeptides derived from high molecular weight kininogen. FASEB J. 2001 November 1, 2001;15(13):2365-76.
16. Passos GF, Medeiros R, Cheng D, Vasilevko V, LaFerla FM, Cribbs DH. The Bradykinin B1 Receptor Regulates Aβ Deposition and Neuroinflammation in Tg-SwDI Mice. Am J Pathol. 2013;182(5):1740-9.
17. Khan M, Liu Y, Khan M, et al. Upregulation of tissue factor in monocytes by cleaved high molecular weight kininogen is dependent on TNF-alpha and IL-1beta. Am J Physiol Heart Circ Physiol. 2010 2010 Feb;298(2):H652-8.

Categorizing the Relationships between Structurally Congruent Concepts from Pairs of Terminologies for Semantic Harmonization

Zhe He, PhD[1], James Geller, PhD[1], Gai Elhanan, MD[2]
[1]New Jersey Institute of Technology, Newark, NJ; [2]Halfpenny Technologies, Blue Bell, PA

Abstract

In this paper, we are using "structurally congruent concepts" in pairs of terminologies to suggest methods for harmonizing the terminologies. Two concepts are structurally congruent if they are children of the same more general concept and parents of the same more specific concept in two different terminologies. We show that structurally congruent concepts can be interpreted in six useful ways, e.g., as new synonyms. All structurally congruent concepts were found for six terminologies from the UMLS, each paired with SNOMED CT. In total, 1384 concept pairs were discovered. Concepts from a sample of 241 pairs were analyzed by a human expert. It was found that 59.3% indicated alternative classifications of the same general concept. This discovery allows an ontology designer to make existing, implicit knowledge explicit. Another 14.5% were newly discovered synonyms, 23.6% suggested the import of a concept into a terminology and 2.5% indicated errors in a terminology.

Introduction

Semantic interoperability is one of the big challenges in biomedical informatics. In order to enrich the semantics and coverage of a terminology and facilitate translational biomedical informatics to be utilized in clinical and research applications, semantic harmonization efforts have recently been extended for various terminologies, e.g. SNOMED CT [1]. However, structural methodologies for semantic harmonization of terminologies have not been studied sufficiently. Weng et al. [2] discussed a conceptual design of a collaborative system for semantic harmonization. Three key design principles were defined: (1) reuse, (2) collaboration, (3) harmonization as modeling. The BRIDG model was presented as a user-centric semantic harmonization framework [3]. The harmonization in the BRIDG model is based on the concept definitions, attributes, and concept relationships. Due to the fact that BRIDG participants are distributed across organizations and no implementation-specific information is provided, it may be hard to use this approach directly by application-oriented users. Tao et al. have discussed the importance of ontology harmonization before using ontologies to annotate clinical data [4]. In this paper, we are approaching semantic harmonization by analyzing the relationships between *structurally congruent* concepts from pairs of terminologies in the UMLS. An outline of the implementation details for finding such structurally congruent pairs is provided.

Auditing of terminologies may uncover problems such as omissions [5]. Previously, we have developed algorithmic and mixed human-computer auditing methods for the UMLS and some of its source terminologies [6, 7]. Auditing may also discover concepts that are synonymous in real life but are coded as different in the UMLS. Occasionally two terminologies in overlapping domains "cut the world at different joints," which makes ontology alignment [8] and ontology integration difficult. In such a situation, the same conceptual knowledge may be classified in (often orthogonal) different ways. We call these "alternative classifications." In this paper, we are describing the use of structural congruency in pairs of terminologies to alert a human auditor to possible cases of harmonization and correction. Due to the importance of SNOMED CT (abbreviated as "SNOMED"), we focus on its concepts.

Background

SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms) [9-11] is considered to be of increasing importance in Medical Informatics. One reason for this status is related to government mandates of using Electronic Health Record systems, meaningful use and incentive payments to physicians. By 2015, SNOMED will become the standard terminology for EHR encoding of diagnoses and problem lists [12]. SNOMED is to be used to "enable a user to electronically record, modify, and retrieve a patient's problem list for longitudinal care (i.e., over multiple office visits)." Thus, in this paper, we are focusing on categorizing the relationships between structurally congruent concepts, one from SNOMED, the other from six reference terminologies. The Unified Medical Language System's (UMLS) [13-16] Metathesaurus [17, 18] is an excellent source of pairs of terminologies with matched concepts. The 2012AB Metathesaurus contains more than 2.8 million concepts and 8.6 million unique concept names from about 160 source vocabularies [19]. SNOMED is also included in the UMLS.

Previously, Bodenreider performed a study of redundant relations and similarity across families of terminologies and discussed the relationship between redundancy and semantic consistency [20]. Bodenreider observed ([21]) that it is

the policy in the UMLS that 'PAR' represents an explicit parent-child relationship in a source, and 'RB' indicates an implied one (as interpreted by the UMLS editorial team). In this paper, we are focusing on explicit hierarchical relationships, thus only terminologies in the UMLS with 'PAR' links annotated with 'IS_A' relationship attributes were chosen. This current work is also marginally related to research on density and granularity of terminologies. Kumar *et al*. [22] lay out a comprehensive theory of granularity in the context of medical terminologies. Schulz *et al.* identify granularity-related problems with "cross-granularity integration" in the biomedical domain [23]. Rector *et al.'s* analysis provides logical formulations of important distinctions between density and related properties [24].

Methods

Our method is based on comparing two medical terminologies from the UMLS. We formally define the targets of our investigation as follows.

Definition: The concepts X (from Terminology 1) and Y (of Terminology 2) are called "structurally congruent" if:

a) Both concepts X and Y have the same parent A in Terminology 1 and in Terminology 2.

b) Both concepts X and Y have the same child B in Terminology 1 and in Terminology 2.

c) The concept X does not appear anywhere in Terminology 2.

d) The concept Y does not appear anywhere in Terminology 1.

e) There is no synonymy relationship and no hierarchical relationship between X and Y (in the UMLS).

Figure 1 shows an abstract layout of two structurally congruent concepts to elucidate the above definition.

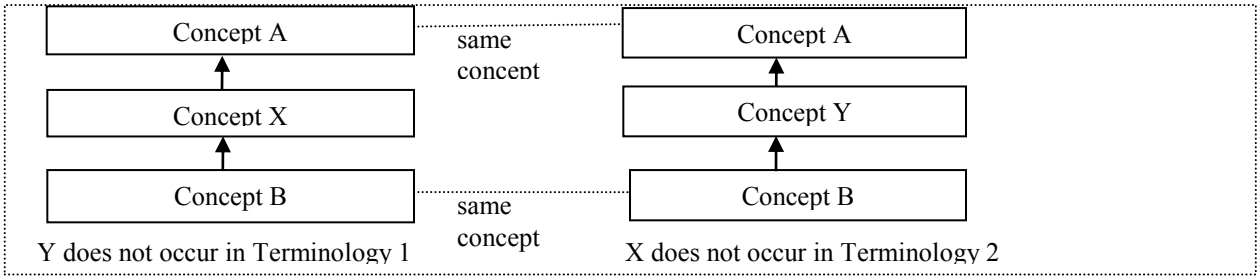


Figure 1. An abstract layout of structurally congruent concepts

It is hypothesized that there are **six possible cases** for how X and Y may relate to each other.

1) The concepts X and Y are alternative classifications. That means that concept A may be validly assigned X and Y as its children. However, these two assignments are indicative of two different ways of clustering the grandchildren of A. Furthermore, concept B may be correctly classified as a child of X and as a child of Y. However, Terminology 1 omits the classification by Y and Terminology 2 omits the classification by X.

2) It holds that B IS_A Y, Y IS_A X, and X IS_A A. In other words, Y may be inserted as a child of X into Terminology 1, thereby adding more detailed information to Terminology 1. Similarly, X may be inserted as a parent of Y into Terminology 2. Such insertions should only be done with approval of a subject matter expert.

3) It holds that B IS_A, X IS_A Y, and Y IS_A A. This is the mirror case of Case 2) in that now X may be inserted as a child of Y into Terminology 2 and Y may be inserted as a parent of X into Terminology 1.

4) Concept X is a real world synonym of concept Y, which was previously not recognized by the UMLS editors.

5) There might be a structural error in Terminology 1, e.g., X is not really a child of A.

6) There might be a structural error in Terminology 2.

Every one of these six cases may be utilized in a human review, possibly leading to an improvement and harmonization of both terminologies. To further probe the potential of this idea, we performed the following study. Six terminologies were selected from the 2012AB release of the UMLS to function as reference terminologies for SNOMED. (Note: It is a *coincidence* that there are six cases and six terminologies.) Only English-language terminologies using the "PAR" relationship annotated with "IS_A" *relationship attributes* were chosen. They are MEDCIN3_2012_07_16, National Cancer Institute Thesaurus (NCI2012_02D), Gene Ontology (GO20 12_04_03),

Medical Entities Dictionary (CPM2003), UMDNS: product category thesaurus (UMD2012) and Foundational Model of Anatomy Ontology (FMA3_1). Due to the fact that the University of Washington Digital Anatomist (UWDA) consists of the Anatomy component and selected structural relationships of FMA, UWDA was excluded even though it also uses "PAR" relationships and "IS_A" relationship attributes. The algorithms were implemented in the Oracle Relational Database Management System (RDBMS) native programming language PL/SQL. The algorithms were used for finding all structurally congruent pairs of concepts, one taken from the list of six reference terminologies, the other one being the July 2012 version of SNOMED. The UMLS is well known to contain many cycles [21, 25], which were eliminated during processing.

Results

Table 1 shows the numbers of pairs of congruent concepts of six reference terminologies relative to SNOMED and the sizes of the samples we randomly chose for human review. The third column shows the number of pairs of congruent concepts found by the program. For reference terminologies with over 100 pairs of congruent concepts, random samples of 70 were chosen for human review; for the others, all of the congruent concepts were reviewed. In total, we reviewed 241 /1384 = 17.4% of all the congruent concept pairs discovered by the program.

Table 1. Comparison of SNOMED CT with six reference terminologies

Reference Terminology	Size of Terminology	# of Pairs of Congruent Concepts	Sample Size
MEDCIN3_2012_07_16	279529	655	70
NCI2012_02D	95523	582	70
FMA3_1	82062	116	70
UMD2012	15956	18	18
GO2012_04_03	61925	6	6
CPM2003	3078	7	7
Total	--	1384	241

The author GE, a medical informaticist and MD with many years of experience in auditing terminologies reviewed the sample. Table 2 shows the results according to the six cases defined in the Methods section. The results show that 59.3% are alternative classifications. Another 14.9% + 8.7% = 23.6% fall into the category where the congruent concept in the reference terminology could be imported into SNOMED, and vice versa.

Table 2. Review results by reference terminology

Reference Terminology	Sample Size	Alternative Classific.	Y IS_A X	X IS_A Y	Error in Trmgy 1	Error in Trmgy 2	Synonym
MEDCIN3_2012_07_16	70	44	10	7	--	1	8
NCI2012_02D	70	38	12	6	--	3	11
GO2012_04_03	6	2	--	4	--	--	--
CPM2003	7	5	--	--	--	--	2
UMD2012	18	9	1	--	--	--	8
FMA3_1	70	45	13	4	2	--	6
Total	241	143	36	21	2	4	35
Percentage	100%	59.3%	14.9%	8.7%	0.8%	1.7%	14.5%

Figure 2 shows an example where congruent concepts were identified as alternative classifications. Thus, *Eleventh posterior intercostal vein* in the FMA is a classification by cardinality, while in SNOMED *Lower right posterior intercostal vein* is a classification by position.

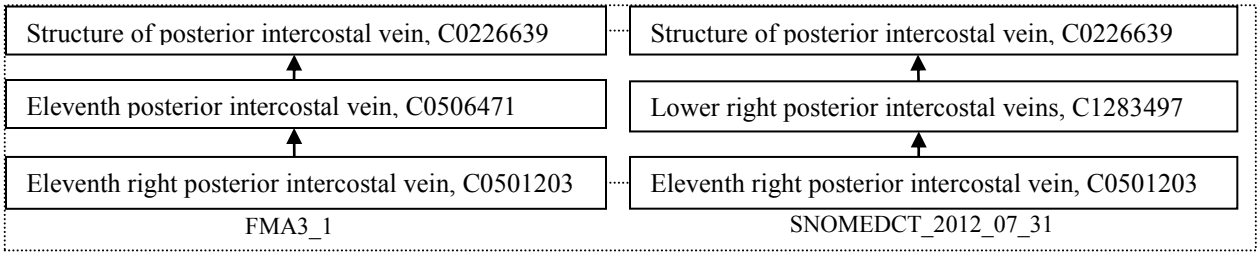


Figure 2. An example of alternative classification

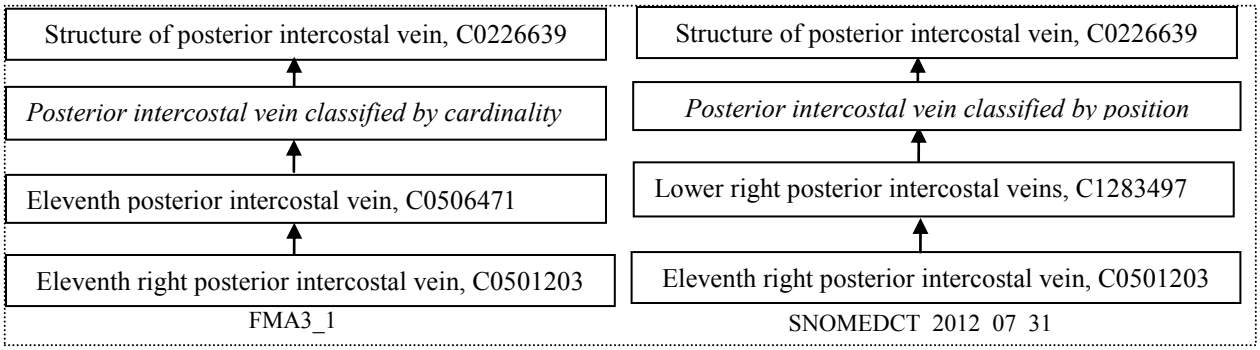


Figure 3. An example of making explicit an implicit assumption of the ontology designers

The discovery of alternative classifications is useful, because it makes explicit the implicit assumptions of the ontology designers how they are viewing the world. This view could then be codified in the ontology. Figure 3 shows the utilization of the findings in Figure 2 by adding two new concepts (with labels shown in *Italics*.)

Figure 4 shows a case where one congruent concept was deemed a parent of the other by the auditor. In this example, the congruent concept *Finding by Site or System* can be a parent of *Finding by site*, thus the congruent concept *Finding by Site or System* from FMA may be added as a parent of *Finding by site* in SNOMED, and vice versa, if this is desirable in the judgment of the owners of the FMA and/or SNOMED.

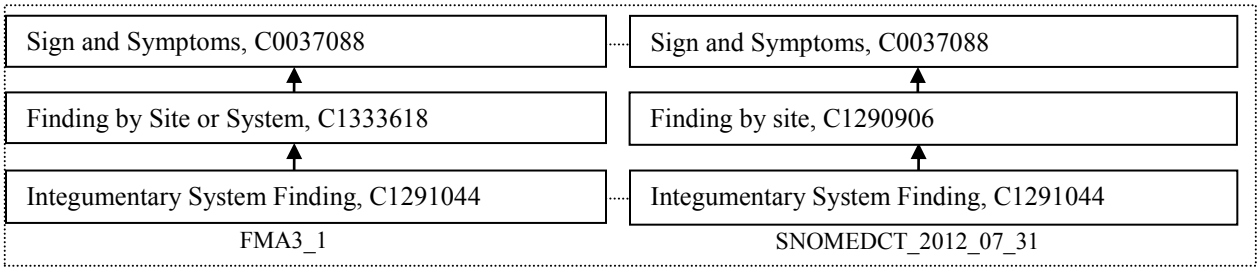


Figure 4. An example of one structurally congruent concept being a parent of the other

The congruent concepts *Chemical Viewed Structurally* from CPM and *Chemical categorized structurally* from SNOMED are deemed synonyms that were not recognized before by our auditor (Figure 5) and should be merged.

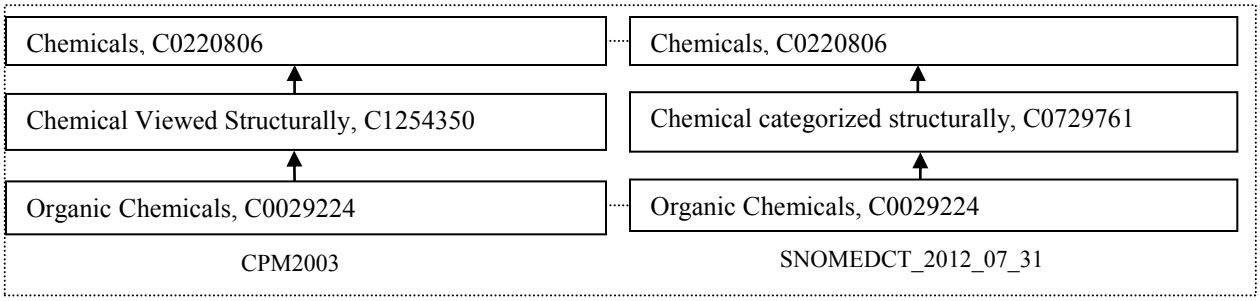


Figure 5. An example of one middle concept being synonymous of the other

During the review of the sample, a few errors within terminologies emerged. The concept from SNOMED *Artificial Implant* was deemed incorrect by the auditor because it should not be considered as "artificial," in the structure with A = Prosthesis, C0175649, Y = Artificial Implants, C0021113, and B = Blood Vessel Prosthesis, C0005846.

Discussion

The UMLS provides many concept pairs from different terminologies, where algorithmically made structural observations raise the question how to harmonize those concepts. In this paper, we identified one such structural observation "structurally congruent concepts" and indicated the different ways how such a congruency can be resolved. However, the semantic harmonization cannot be done without the consent of terminology curators. Moreover, modeling differences between terminologies make semantic harmonization difficult. For UMD2012 (Table 2), eight pairs of congruent concepts were found to be synonyms. For GO, more cases where one congruent concept is a potential parent of the other were found than alternative classifications. For our cases 2) and 3), relevant work in MIREOT [26] defines a set of guidelines for importing classes from external ontologies and proposes an automated mechanism and a minimal information standard for selectively importing classes into an ontology. However, it only supports OBO foundry ontologies (OWL format). In this paper, all the terminologies are in UMLS RRF format. Thus, the import guidelines introduced in MIREOT cannot be used here directly.

A possible limitation of this work is that it uses SNOMED concepts and all reference terminology concepts in the formats that they were provided in by the UMLS. There may be differences between the original concept representation of SNOMED (or the reference terminologies) and the representation of SNOMED that is accessible through the UMLS.

Conclusions and Future Work

Six terminologies of the UMLS were compared with SNOMED with respect to structurally congruent concepts. In a sample study it was found that the great majority of cases corresponded to alternative analysis situations (143 out of 241, corresponding to 59.3%). The second most common situation indicated the possibility of adding more detail to SNOMED CT or the reference terminologies (57 out of 241, corresponding to 23.6%). In 35 cases new synonyms were discovered, and three pairs of concepts indicated errors. As future work, we plan to conduct a study to analyze structurally congruent concepts between pairs of any two META terminologies with explicitly defined hierarchical relationships, e.g., not limited to SNOMED CT being Terminology 2. We are also planning a more extensive evaluation of the results. The work in this paper was limited to pairs of structurally congruent concepts. However, we have noticed cases of congruency that involve three, four and even more concepts. An analysis of these cases is under way.

References

1. IHTSDO. SNOMED CT and LOINC to be linked by cooperative work. 2013 [cited September 29, 2013]; Available from: http://www.ihtsdo.org/about-ihtsdo/governance-and-advisory/harmonization/loinc/
2. Weng C, Fridsma DB. A call for collaborative semantics harmonization. AMIA Annu Symp Proc. Washington D.C.; 2006.
3. Weng C, Gennari JH, Fridsma DB. User-centered semantic harmonization: a case study. J Biomed Inform. 2007 Jun;40(3):353-64.
4. Tao C, Solbrig HR, Chute CG. CNTRO 2.0: A Harmonized Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives. AMIA Summits Transl Sci Proc. 2011;2011:64-8.
5. Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. J Biomed Inform. 2009 Jun;42(3):407-11.
6. Gu H, Perl Y, Geller J, Halper M, Liu LM, Cimino JJ. Representing the UMLS as an object-oriented database: modeling issues and advantages. J Am Med Inform Assoc. 2000 Jan-Feb;7(1):66-80.
7. Chen Y, Gu HH, Perl Y, Geller J. Structural group-based auditing of missing hierarchical relationships in UMLS. J Biomed Inform. 2009 Jun;42(3):452-67.
8. Shvaiko P, Euzenat J. Ontology Matching: State of the Art and Future Challenge. Knowledge and Data Engineering, IEEE Transactions on. 2013;25(1):158-76.
9. Wilcke JR, Green JM, Spackman KA, et al. Concerning SNOMED-CT content for public health case reports. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):613; author reply -4.
10. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp. 2001:662-6.
11. SNOMED CT Homepage. [cited January 10, 2013]; Available from: http://www.ihtsdo.org
12. US Department of Health and Human Services, Health Information Technology: Initial Set of Standards, Implementation Specifications, and Certification Criteria for Electronic Health Records Technology. [cited May 21, 2013]; Available from: http://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17210.pdf

13. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
14. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.
15. Campbell KE, Oliver DE, Shortliffe EH. The Unified Medical Language System: toward a collaborative approach for solving terminologic problems. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):12-6.
16. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc. 1998 Jan-Feb;5(1):1-11.
17. Schuyler PL, Hole WT, Tuttle MS, Sherertz DD. The UMLS Metathesaurus: representing different views of biomedical concepts. Bull Med Libr Assoc. 1993 Apr;81(2):217-22.
18. Tuttle M, Sherertz DD, M. E, Olson N, Nelson S. Implementing Meta-1: The First Version of the UMLS Metathesaurus. Proc Annu Symp Comput Appl Med Care; 1989. p. 483-7.
19. Resource Description Framework (RDF). [cited March 3, 2013]; Available from: http://www.w3.org/RDF/
20. Bodenreider O. Strength in numbers: exploring redundancy in hierarchical relations across biomedical terminologies. AMIA Annu Symp Proc. 2003:101-5.
21. Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp. 2001:57-61.
22. Kumar A, Smith B, Novotny DD. Biomedical informatics and granularity. Comp Funct Genomics. 2004;5(6-7):501-8.
23. Schulz S, Boeker M, Stenzhorn H. How Granularity Issues Concern Biomedical Ontology Integration. In Proceedings of the International Congress of the European Federation for Medical Informatics (MIE 2008). Gothenburg, Sweden; 2008. p. 863-68.
24. Rector A, Rogers J, Bittner T. Granularity, scale and collectivity: when size does and does not matter. J Biomed Inform. 2006 Jun;39(3):333-49.
25. Mougin F, Bodenreider O. Approaches to eliminating cycles in the UMLS Metathesaurus: naive vs. formal. AMIA Annu Symp Proc. 2005:550-4.
26. Courtot M, Gibson F, Lister AL, Malone J. MIREOT: The Minimum Information to Reference an External Ontology Term. In: Smith B, editor. International Conference on Biomedical Ontology. Buffalo, New York, USA; 2009. p. 87-90.

Automated, Quantitative Analysis of Histopathological Staining in Nuclei

Ricardo Henao[1], PhD, Joseph Geradts[1], MD, MA, Manabu Kurokawa[2], PhD, Sally Kornbluth[1], PhD, and Joseph E. Lucas[3], PhD
[1]Duke University, Durham, NC
[2]Dartmouth College, Hanover, NH
[3]Quintiles, Durham, NC

Abstract

Technological advances have allowed the generation of high-throughput imaging of tissue sections. However, the analysis of these samples is typically still performed manually by one or multiple pathologists. We present a novel statistical model for the automated, quantitative analysis of these images. Our approach requires minimal tuning and allows recapitulation of estimates of staining strength in the nuclei of tumor cells as estimated by the gold standard. Besides, it compares favorably to other quantitative approaches available in the public domain.

Introduction

Analysis of tissue sections after staining is a subjective and labor intensive process. Typically, the pathologist must manually scan through a series of slides to estimate the strength of staining using a discrete scoring system. Often only a subset of cells on the slide should be considered and often in only one region of those cells. For large studies, this process may involve multiple pathologists, which leads to challenges with subjectivity, inter-rater reliability, and fatigue. In this paper we describe a model for the automated analysis of stained slides that results in an objective, repeatable and quantitative assessment of the staining level of a particular protein in the nuclei of cancer cells.

There are some targeted approaches to this task in the literature, such as Masseroli et al (2000)[1] for liver fibrosis and Davis et al (2003)[2] for apoptosis. The most rudimentary general approach is to simply integrate the total amount of staining of the relevant color across a slide. However, this does not allow for assessment of stain levels in different parts of the slide separately. One approach to addressing this problem involves the manual selection of relevant regions together with image manipulation by the pathologist[3]. This leads to a quantitative result, but is still not tractable in a high-throughput experiment. This challenge has inspired the development of some excellent toolkits to support automation of this task[4,5].

Dirichlet mixture models have been successfully used in a wide range of image processing applications. More particularly, various specialized Bayesian models for segmentation tasks based on Dirichlet processes have been proposed in recent years[6,7]. However, most of this work targets natural images. The development of segmentation models, such as the one described in this paper, employing Dirichlet processes appears to be very promising for the analysis of histopathological images.

In this paper we present a statistical model for the automated analysis of histopathology images of tumor sections that is able to recapitulate the pathologist's assessment in a quantitative and repeatable way. Our model requires very little tuning and compares favorably to other publicly available software designed for this task.

Data

The data set consists of 30 digitized microscope images. Each image is RGB encoded, 200x magnification scale and 1100×828 pixels in size. All shots were manually taken and subsequently labeled by a pathologist

into either low or high levels of expression of the stained protein. Figure 3 show particular examples of images of these two classes.

Our approach employs two steps. We start by extracting some features from the images, specifically we over-segment each image into small patches known as superpixels. These are commonly used for image labeling problems in which pixel-level labeling might be prohibitive or as preprocessing step in more complex segmentation algorithms. We use it as a way to obtain a compressed color representation of images. The main goal in this step is to isolate nuclei of tumor cells from background. This serves two purposes: (i) eases visualization by shifting focus to nuclei and (ii) information gathered from segmented nuclei can be used for further analysis or as part of a more elaborated pipeline. We employed turbopixels, a fast superpixel algorithm based on geometric flows[8]. For each superpixel we compute a 128 bin RBG histogram, resulting into a 384 dimensional vector containing binned color counts. We did not notice significant changes in results by further increasing the number of bins used to compute histograms. Provided that each image was divided into approximately 8600 superpixels, each image is then represented as a 8600×384 integer matrix. The second step of our analysis pipeline involves the fitting of a hierarchical statistical model to the resulting set of 30 matrices.

Model

The Dirichlet process (DP) has been widely used to dynamically model the number of clusters in conjunction with mixture models. The Chinese Restaurant Process[9] (CRP) offers a very useful metaphor for the DP and some of its generalizations[10]. Imagine a restaurant with an infinite number of tables denoted by $\boldsymbol{\phi}_k$, customers $\boldsymbol{\theta}_n$ enter sequentially to the restaurant so that the n-th customer sits at a given table with probability proportional to the number of customers already occupying it m_k or gets a new table with probability proportional to α. From the metaphor we can see that the DP has a rich-get-richer dynamic and that α, the concentration parameter, controls the total number of tables (clusters) for a given customer base. The CRP results in an exchangeable model, i.e. the probability distribution over partitions does not depend on the ordering of the customers. Defining K as the number of non-empty tables and $z_n \in 1, \ldots, K$ to be the table assignment for customer n, the prediction rule in the CRP can be written as

$$z_n | \mathbf{z}_{\setminus n}, \alpha \;\propto\; \alpha \delta_{k^\star} + \sum_{k=1}^{K} n_k \delta_k \,,$$

where $\mathbf{z}_{\setminus n}$ is the set of table assignments excluding customer n and $k^\star$ indexes a new table. In terms of our microscope images we can think of superpixels as customers, $\boldsymbol{\phi}_k$ as color profiles or color distributions and z_n as to which color profile superpixel $\boldsymbol{\theta}_n$ belongs to. More formally we can say that $\boldsymbol{\theta}_n$ is a sample from a distribution $F(\boldsymbol{\phi}_{z_n})$ with parameter $\boldsymbol{\phi}_{z_n}$, and $\boldsymbol{\phi}_k$ follows a DP with concentration parameter α and base measure H. Provided that superpixels are represented as quantized color histograms, we assume a discrete distribution for $F(\boldsymbol{\phi})$ and let H be a Dirichlet distribution with concentration parameter γ.

We want to obtain a compact color based representation of each image, however we still want to be able to share information across them. In principle, we can add a top layer to the CRP model to enable color profile sharing, i.e. each image will have its own CRP (bottom layer) which in turn will get its color profiles from a common CRP. This model is commonly known as the Chinese restaurant Franchise (CRF) representation of the hierarchical Dirichlet process (HDP)[11]. Figure 1 shows a graphical representation of our proposed HDP model for microscope images. The top layer of the model is represented by two parameters $\boldsymbol{\phi}_k$ and β_k, being a 384 dimensional normalized color profile and its probability of occurrence, respectively. The potentially infinite set of color profiles is indexed by k, meaning that in practice we only instantiate color profiles for which $\beta_k > 0$. Hyperparameters α_0 and γ control the total number of *active* color profiles and their spectrum *specificity*, respectively. For instance if γ is very small, profiles will encode very narrow wavelength bands. The bottom layer has the image specific parameters; $\boldsymbol{\theta}_{ni}$ is color histogram for superpixel n in image i, z_{in} is the assignment variable for superpixel n in image i, $\boldsymbol{\pi}_j$ is the vector of color profile usage frequencies for image i and N_i is the number of superpixels in image i. Lastly, hyperparameter α controls the total number of color profiles used by each image.

The model has various parameters of interest namely z_{in}, $\boldsymbol{\phi}_k$ and $\boldsymbol{\pi}_i$. Inference is carried out using Markov Chain Monte Carlo (MCMC) and hyperparameters $\{\alpha, \alpha_0, \gamma\}$ were provided with prior distributions

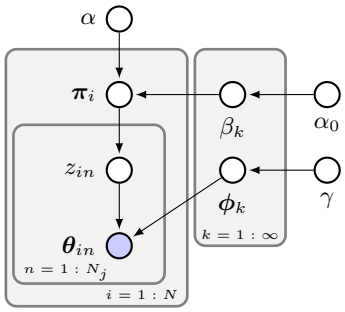


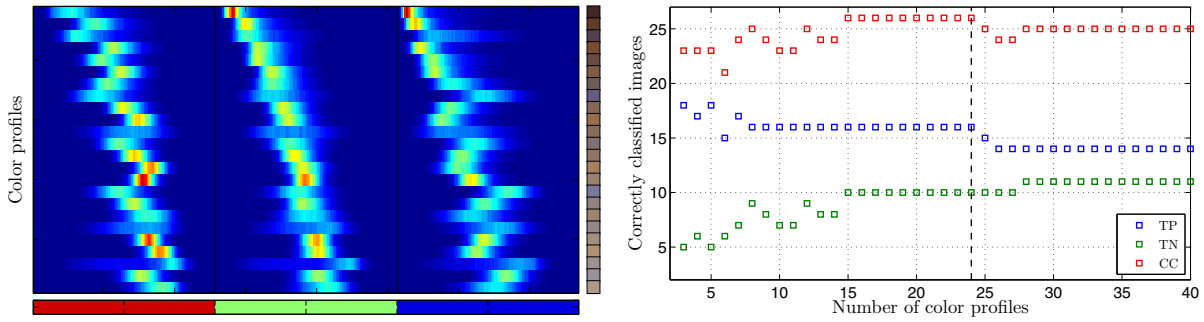
Figure 1: Graphical model for the HDP model. N is the number of images, $\{\alpha, \alpha_0, \gamma\}$ is the set of hyperparameters and $\boldsymbol{\theta}_{in}$ is the only observed variable in the model (shaded node). We used bold letters to distinguish vectors from scalars.

to facilitate their tuning. In particular, gamma priors with shape 2 and rate 1 where used everywhere. There are several MCMC sampling approaches for HDPs, we are using a truncated DP representations with a maximum of 200 color profiles. In practice we did not observe the model reaching the profile limit at any time during inference, meaning that further increasing the maximum number of color profiles does not change the results. Inference details can be found for instance in Teh et al (2006)[11].

Results

The previously described superpixel processed data consist of a 260457×384 matrix of quantized color histograms. The HDP sampler was run for 1500 iterations, we observed in general good mixing and cluster assignment stabilization after the first 500 iterations. We did not notice significant changes in the cluster assignments after making small changes in all the hyperparameters settings of the model, i.e. shapes and rates of gamma hyper-priors. We tried to make quantitative comparison of the proposed HDP model against well established algorithms for image segmentation including normalized cuts[12] (Ncuts), mean shift[13] (MS) and K-means in conjunction with superpixels. Ncuts was computationally prohibitive considering the size of the images in the data set. We tried to select for K in K-means using internal measures such as Davies-Bouldin index[14] and silhouette average[15] but segmentation results were too coarse for the desired level of detail. For MS we observe nice segmentation results when manually tuning its parameters however we could not find a good set of parameters by grid search and internal measures to appropriately fit the entire data set. It is still possible that these methods could work satisfactorily with additional preprocessing or specialized parameter tuning.

After running the HDP inference we ended up with a model with 180 color profiles. We know that nuclei appear darker than the remaining elements of the background thus we can simply sort color profiles according to intensity to then set a manually selected threshold for visualization purposes. Besides, just being able to interactively set the threshold could be a very useful tool for exploratory purposes. Here we attempt to select the *best* number of color profiles according to their ability to correctly classify the status of each image. In order to do so in an unbiassed manner, we perform leave-one-out cross-validation (LOOCV) using a number of color profiles ranging from 2 to 180 and a naive Bayes classifier model[16]. At stage j of LOOCV, classifier training is done using a subset of the color profile usage probabilities $\boldsymbol{\pi}_i$ for all images but the one being tested, then HDP model and trained classifier are used in turn to make a prediction. Figure 2(b) shows classification results for a range of color profiles, we see that the accuracy curve is rather flat meaning that the classifier does a good job for a set of color profiles ranging from 15 to 24. This also indicates that selecting the number of profiles for discrimination purposes is not a critical task in our case. For visualization purposes we select the number of profiles in order to maximize classification accuracy, i.e. 24 color profiles. We can also see from Figure 2(b) that at 86.7% accuracy (26/30 images), true positive and true negative rates are 88.9% (16/18) and 83.3% (10/12), respectively, which suggests that the classifier has a well balanced misclassification risk. We did not consider more sophisticated classifiers, however we believe that classification accuracy can be further improved with an upgraded classifier or better yet by integrating it directly into the HDP model. If we examine the 24 selected color profiles in Figure 2(a) we see that most



(a) (b)

Figure 2: Classification results. (a) The 24 color profiles used for final image segmentation, there were obtained from the LOOCV procedure. Blocks in the left bar show the mean color encoded by each profile. The vertical lines separate individual sections of RGB spectrum. (b) LOOCV number of correctly classified (CC), true positives (TP) and true negatives (TN) images using a naive a Bayes classifier. The vertical dashed line denotes the selected number of color profiles.

of them summarize the desired color features, i.e. dark/middle brown and bluish shades characteristic of nuclei in our image data set.

Figure 3 shows examples of segmented images one from each category, i.e. low and high levels of expression of the stained protein. We see that the HDP based segmentation model with classifier aided profile selection produces a very nice separation between background and cell nuclei despite color heterogeneity. Larger versions of the images in the figure can be found online at: `http://people.duke.edu/~rh137/huwe1.html`.

Closing remarks

We foresee a version of the presented model in which the classifier is integrated into the segmentation model as an additional layer, in this way the model will be able to bias color profile assignments towards better classification performance and hopefully improve visual representations of nuclei data.

One particularity of the data we have not addressed yet but represents a good opportunity for overall improvement is to extend the model to use morphological information about the segments/superpixels, for instance size or regularity. We know that this kind of information is used by pathologists to better inform their decisions.

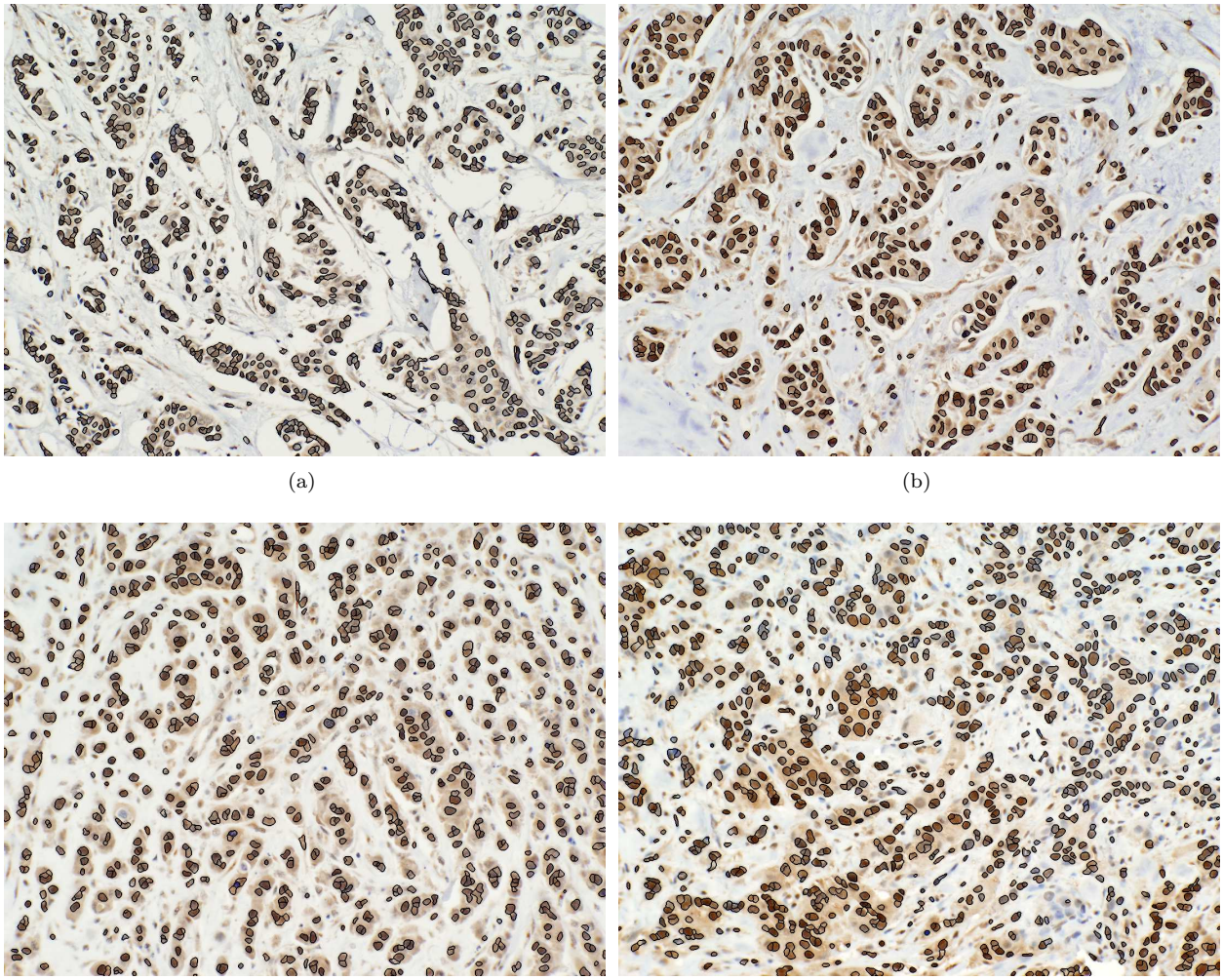
References

[1] Masseroli M, Caballero T, O'Valle F, Moral RMGD, Pérez-Milena A, Moral RGD. Automatic quantification of liver fibrosis: design and validation of a new image analysis method: comparison with semi-quantitative indexes of fibrosis. Journal of hepatology. 2000;32(3):453–464.

[2] Davis DW, Buchholz TA, Hess KR, Sahin AA, Valero V, McConkey DJ. Automated Quantification of Apoptosis after Neoadjuvant Chemotherapy for Breast Cancer Early Assessment Predicts Clinical Response. Clinical cancer research. 2003;9(3):955–960.

[3] Lehr HA, Jacobs TW, Yaziji H, Schnitt SJ, Gown AM. Quantitative evaluation of HER-2/neu status in breast cancer by fluorescence in situ hybridization and by immunohistochemistry with image analysis. American journal of clinical pathology. 2001;115(6):814–822.

[4] Peng H. Bioimage informatics: a new area of engineering biology. Bioinformatics. 2008;24(17):1827–1836.



(a) (b)

(c) (d)

Figure 3: Segmentation examples. Black lines are introduced to highlight nuclei segments corresponding pixels assigned to one of the 24 color profiles selected by the classification model. Images correspond to low (a and c) or high (b and d) levels of expression of the stained protein.

[5] Roysam B, Lin G, Bjornsson C, Narayanaswamy A, Chen Y, Shaina W, et al. The FARSIGHT project: associative multi-dimensional image analysis methods for optical microscopy. In: J Rittscher SW R Machiraju, editor. Microscopic Image Analysis for Life Science Applications. Artech Publishing House; 2008. .

[6] Du L, Ren L, Dunson D, Carin L. A bayesian model for simultaneous image clustering, annotation and object segmentation. Advances in Neural Information Processing Systems. 2009;22:486–494.

[7] Ghosh S, Ungureanu AB, Sudderth EB, Blei DM. Spatial distance dependent Chinese restaurant processes for image segmentation. In: Shawe-Taylor J, Zemel RS, Bartlett P, Pereira FCN, Weinberger KQ, editors. Advances in Neural Information Processing Systems 24. MIT Press; 2011. p. 1476–1484.

[8] Levinshtein A, Stere A, Kutulakos KN, Fleet DJ, Dickinson SJ, Siddiqi K. Turbopixels: Fast superpixels using geometric flows. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2009;31(12):2290–2297.

[9] Aldous D. Exchangeability and related topics. École d'Été de Probabilités de Saint-Flour XIII—1983. 1985;p. 1–198.

[10] Pitman J. Combinatorial stochastic processes. vol. 1875 of Lecture notes in mathematics, Ecole d'ete de probabilités de Saint-Flour XXXII. Berlin: Springer-Verlag; 2006.

[11] Teh YW, Jordan MI, Beal MJ, Blei DM. Hierarchical Dirichlet processes. Journal of the American Statistical Association. 2006;101(476):1566–1581.

[12] Shi J, Malik J. Normalized cuts and image segmentation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2000;22(8):888–905.

[13] Comaniciu D, Meer P. Mean shift: A robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2002;24(5):603–619.

[14] Davies DL, Bouldin DW. A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1979;(2):224–227.

[15] Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics. 1987;20:53–65.

[16] Bishop CM. Pattern Recognition and Machine Learning. Springer; 2006.

Automated Tools for Clinical Research Data Quality Control using NCI Common Data Elements

Cody L. Hudson, Umit Topaloglu, PhD, Jiang Bian, PhD, William Hogan, MD, Thomas Kieber-Emmons, PhD,
University of Arkansas for Medical Sciences, Little Rock, AR

Abstract

Clinical research data generated by a federation of collection mechanisms and systems often produces highly dissimilar data with varying quality. Poor data quality can result in the inefficient use of research data or can even require the repetition of the performed studies, a costly process. This work presents two tools for improving data quality of clinical research data relying on the National Cancer Institute's Common Data Elements as a standard representation of possible questions and data elements to A: automatically suggest CDE annotations for already collected data based on semantic and syntactic analysis utilizing the Unified Medical Language System (UMLS) Terminology Services' Metathesaurus and B: annotate and constrain new clinical research questions though a simple-to-use "CDE Browser." In this work, these tools are built and tested on the open-source LimeSurvey software and research data analyzed and identified to contain various data quality issues captured by the Comprehensive Research Informatics Suite (CRIS) at the University of Arkansas for Medical Sciences.

Introduction

With emerging healthcare technologies and systems becoming increasingly reliant on the efficient and expedient transfer of data between disparate systems, the assessment and maintenance of data quality of healthcare and clinical data has become prominent areas of research and effort in the electronic healthcare frontier[1]. Though many standards, technologies, and vocabularies exist to aid in supplying a maintainable level of data quality in healthcare systems, such as the HL7 messaging standard[2], caCORE[3], or the Unified Medical Language System[4], there still exists significant hurdles and inadequacies in current methods for ensuring high data quality[5, 6, 7, 8]. As a facet of the entire quality problem presented by healthcare and clinical data, this work focuses on the data quality of clinical research data, describing and implementing two tools that utilize the National Cancer Institute's (NCI) Common Data Elements (CDE)[9] as a syntactic standard and the vocabulary accessed through the UMLS Terminology Service's (UTS) Metathesaurus as a semantic standard for automated syntactic/semantic annotation of past clinical research data and as a library for computer-aided syntactic/semantic annotation and constraint of new clinical research questions for future studies. Through the supplied annotations, it is the hope of this work that data quality can be improved between disparate sources of clinical research through means of a standard semantic and syntactic representation of any and all produced research data as well as ensured data quality through enforced syntactic/semantic constraints. To explore the effectiveness of the proposed approach in achieving the aforementioned goals, two tools were developed, noted respectively as the Automated Annotation Tool (which annotates questions with CDEs based on minimizing semantic and syntactic distance between survey questions and potential CDEs) and the CDE Importer (a plugin for Limesurvey forcing users to annotate questions with CDE codes). These tools were implemented using the LimeSurvey software as a basis for clinical research data collection and run against clinical research data generated by the University of Arkansas for Medical Sciences (UAMS)'s Comprehensive Research Informatics Suite (CRIS). Here we describe our implementation focusing on relevant information for LimeSurvey as a clinical research tool, information concerning CDEs as defined by NCI, background information concerning UMLS, UTS, and the Metathesaurus, the methodology employed by the two tools (Automated Annotation Tool and CDE Importer) implemented in this work. We provide sample results of using the Automated Annotation Tool on live clinical research survey data captured with LimeSurvey, and final remarks detailing planned improvements on the Automated Annotation Tool and CDE Importer.

Background

Data Quality *Processes*

The Ten Step Process[10] is used to assess, improve, and create high-quality information with long-lasting results. It should be considered an evolving and continuous process to improve the quality of the data with the following phases;

- The Assessment Phase – this phase includes identifying business needs, analyzing information environment and conducting data quality assessment.

- The Awareness Phase – studying the root causes of data quality problems identified in assessment phase.
- The Action Phase –Implementing plans which are developed at the awareness phase.

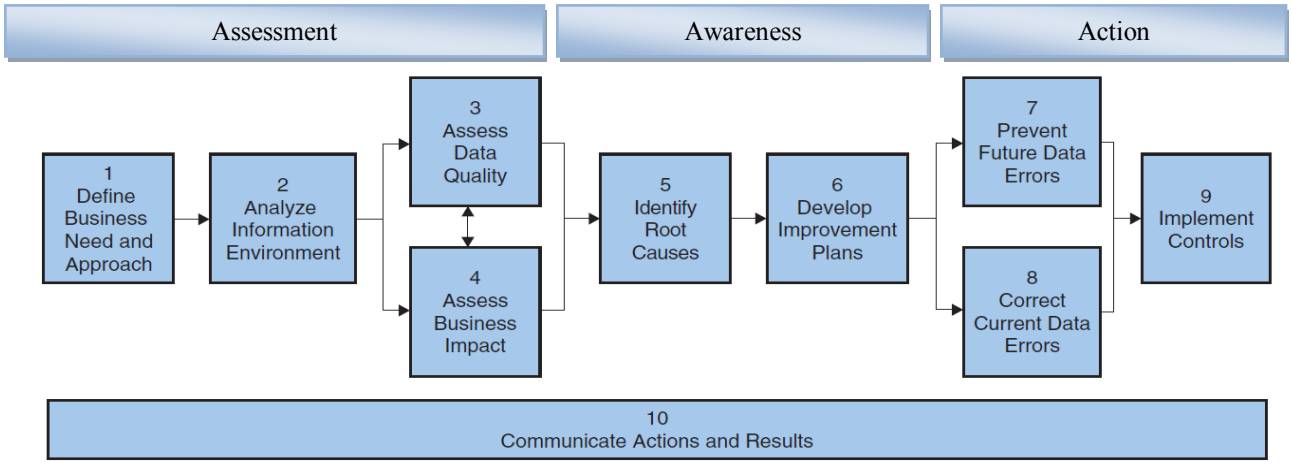


Figure 1. The Information and Data Quality Improvement Cycle and the Ten Steps Process[10]

In the context of data quality improvement cycle, the following phases were examined: "Assessment of actual environment"; "Awareness to understand true state of data and their impact on business"; and "Action to prevent future information quality problems and correction of existing problems." These phases are essentially the main components of any improvement life cycle that can be used by individuals or teams alike. The improvement life cycle is the basis for the Ten-Step Process that provides explicit and detailed instructions for planning and executing quality improvement projects by combining data quality dimensions and business impact techniques.

Getting to know the data

The core of the data utilized in this work originates with several research databases collected over the years by the Cancer Control Program of the Winthrop P. Rockefeller Cancer Institute. As initial assessment phase of the Ten Step Process, a data profiling process was implemented on the provided data and a number of data quality issues were identified[15]. Some of the main issues identified were:

- Information obtained via an existing process generated difficult-to-classify values due to lack of standardization, consistency, and commonly accepted data elements.
- Lack of clarity or absence of acceptable forms of responses.
- Lack of data collection mechanisms to enforce constraints and quality controls.
- Accuracy problems (i.e. incorrect values).
- Completeness problems.
- Data pattern problems.
- Duplication problems.

The quality issues that were identified using this Ten Step Process were found to be addressable in either a manual or automated manner. Manual processes have been utilized in past works to address issues related to duplicate records (e.g. duplicate participant IDs). Conversely, this work provides automation for defining standardized annotations via CDEs, given that manual CDE annotation can be costly both in terms of time, money, and the required training to properly utilize CDEs.

UAMS CRIS

The Comprehensive Research Informatics Suite, or CRIS (formerly known as the Clinical Trials Informatics Suite) is a software suite developed and packaged by UAMS for distributed electronic maintenance and deployment of

clinical trials and all related data. The suite provides functionality for subject registration, research study calendars and patient activities management, automated coding using standard medical vocabularies for supplied free text, electronic participant recruitment for clinical trials, date tracking, data reporting, and electronic data capture using tools such as OpenClinica and LimeSurvey. The tools developed in this work are built to utilize and annotate the data captured through CRIS, and all test data used in this work was captured using the CRIS system.

LimeSurvey

LimeSurvey is one of the primary applications for capturing clinical research data using UAMS's CRIS. This open source, free survey software provides an extremely flexible platform and wide host of tools for developing surveys and survey questions through an intuitive interface. More notably, LimeSurvey offers excellent tools for constraining the syntax of possible answers that can be provided by survey users, such as the base question type (e.g. string, numeric, multiple choice, date, etc.) or through constrains such as minimum/maximum characters allowed, minimum/maximum values allowed, as well as user-defined regular expressions.

NCI Common Data Elements

Common Data Elements, or CDEs, are standardized metadata constructs that can describe both the syntactic and semantic constraints of an entity, such as a patient name or a street address. The National Cancer Institute (NCI) developed CDEs specifically for cancer research (though it now includes a wide variety of contexts) to address the data control issue present with the creation of new, dissimilar data elements per individual researcher. The main resource for accessing the all NCI CDEs is through the Cancer Data Standards Repository, or caDSR, which offers available web services such as the REST API for programmatically querying caDSR and retrieving CDEs. Manual queries can be performed with NCI's CDE Browser (not to be confused with the CDE Importer implemented and explained in this work).

Each NCI CDE has various attributes that make it particularly useful for the goals of this work. At the basis of each CDE is a "data element" that describes top level attributes such as a unique identifier, a preferred name, preferred question text (when utilized in clinical trials research), a workflow status (of being integrated into caDSR), all relevant contexts, any previous versions of the CDE, as well as other related information. Semantic information is captured in terms of the "data element concept," providing unique codes that link to concepts defined in NCI's Thesaurus, specifically the CDE's "Object" and "Property" codes describing, in turn, the real world entity and attributes described by a given CDE. Each CDE also contains a "value domain" describing the syntactic constraints defined by the CDE, such as the data type, minimum/maximum character length, minimum/maximum values allowed, any permissible values for enumerated value domains, etc. With both the "data element concept" and "value domain," each CDE contains sufficient metadata to describe a concept both syntactically and semantically. For this reason, CDEs were chosen in this work to annotate clinical research data and apply constraints on new clinical studies. Furthermore, despite shortcomings in the CDE database and design, current studies suggest that CDEs are an effective tool for providing data quality assurance[11, 12, 13].

Unified Medical Language System

The Unified Medical Language System, or UMLS, provides access to a large number of cross-referenced vocabularies that describe concepts semantically through relational mappings and semantic metadata. Two prominent components of UMLS include the Metathesaurus and Semantic Network, each which can be accessed and queried either manually or programmatically through the UMLS Terminology Services (UTS) by authorized users. The Metathesaurus provides access to "concepts" that contain relational mappings to an enormous number of vocabularies, such as SNOMED CT, RxNorm, and, most prominent to this work, the NCI Thesaurus. Each concept is defined by a preferred name, "atoms" which represent mapped concepts present in other vocabularies with their respective relations, and one or more semantic types. The semantic types map each concept to other concepts that share the same semantic type or are defined by a related semantic type as specified by the aforementioned Semantic Network. The Semantic Network is organized in a tree hierarchy, with more general semantic types defined at the root. Each semantic type, just as with the concepts, contains a list of related semantic types and their respective relations, as well as other auxiliary information. As UMLS contains semantic mappings to concepts defined in the NCI Thesaurus, it is utilized in this work to convert Concept Unique Identifiers (CUIs) extracted from question text using the MetaMap[16] utility into NCI Thesaurus codes.

Methods

Automated Annotation Tool

To provide syntactic/semantic annotation of past clinical survey data, the Automated Annotation Tool (AAT) was developed to automatically provide CDE annotations for clinical data with very little required human interaction. Given this, the implemented tool was designed to accept a valid LimeSurvey survey table and generate suggested CDE annotations based on minimizing semantic and syntactic distance between each survey question and their respective suggested CDEs. The process for generating the semantic and syntactic distance for that exists between a LimeSurvey question and a potential list of CDEs and thus determining the best CDE (in terms of minimum distance) is expressed in Figure 2 below:

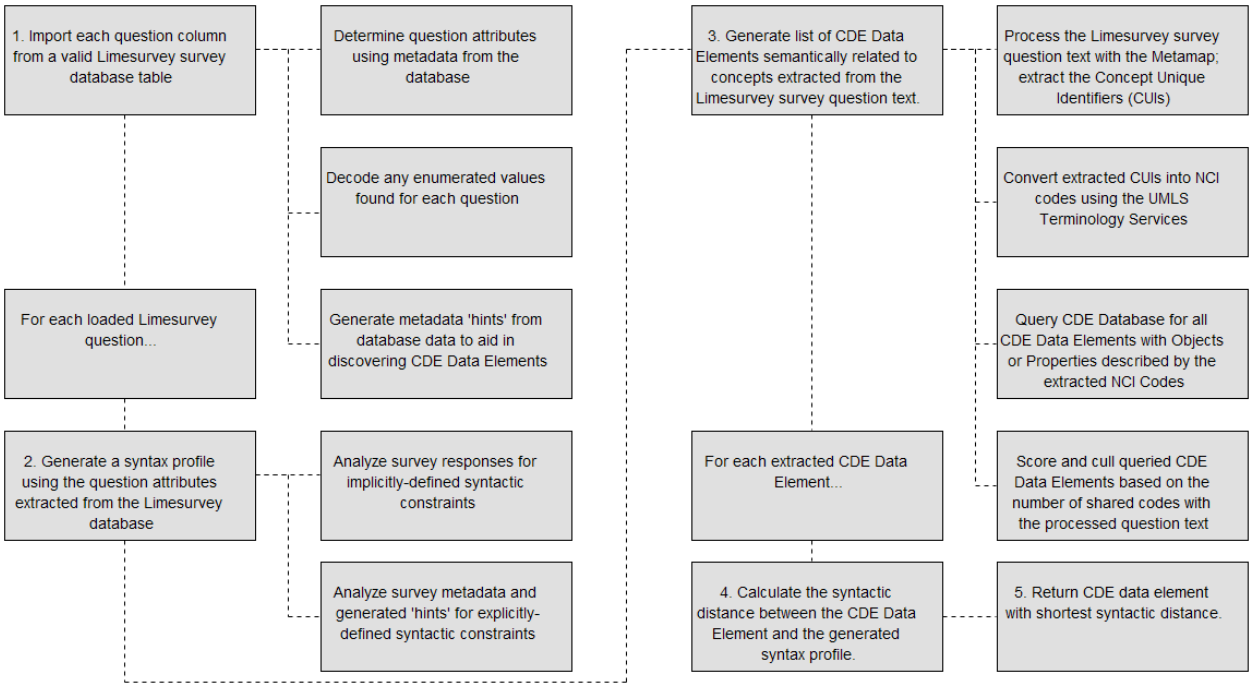


Figure 2. CDE-Question Distance Calculation Process

As shown in Figure 2, there are four primary steps taken to calculate the semantic and syntactic distance that exists between a given question and CDE data element. The first step encompasses extracting the questions and all of their relevant attributes found in the provided LimeSurvey table and database. This process includes determining the question text, question code, the LimeSurvey question type, and any user-specified constraints on the data (such as maximum character length). If the question contains enumerated values as answers, this step then also includes decoding the values and merging any shared enumeration lists. When available, certain metadata attributes can also be used to generate "hints" for the annotation process. One such hint can be generated with the LimeSurvey survey question type; given that many of the CDE data types can be mapped and directly compared to various LimeSurvey survey question types. While such hints allow the AAT to more accurately determine the syntactic distance that exists between a LimeSurvey survey question and a CDE, they are not necessary to perform the distance calculation.

After each question is extracted, the AAT then proceeds to generate a syntax profile utilizing the extracted attributes from step 1 in Figure 2. All extracted explicitly-defined constraints (i.e. defined in the survey metadata) are loaded into the syntax profile; all other implicitly-defined constraints are determined through statistical analysis of the answers provided for each survey question. Constraints specified in a fully initialized syntax profile include the base data type (string, numerical, and enumerated) mapped from the LimeSurvey question type, minimum and maximum character lengths, minimum and maximum numerical precision, minimum and maximum numerical value,

permissible answers for enumerated data types, and finally any available metadata hints. A single syntax profile is generated per survey question; any syntax profiles representing distributed LimeSurvey questions, such as arrays, are merged after the extraction process to increase annotation accuracy and reduce redundant question analysis.

Using the question text contained in the syntax profile for each LimeSurvey question, the third step in the AAT process attempts to implicitly minimize semantic distance by querying CDEs such that the returned set only contains data elements that share one or more semantic concepts contained within the question text. It is from this returned set that the final CDE, the chosen annotation, is determined. To create the set of semantically-related CDEs, a list of Concept Unique Identifiers (CUIs) is first generated from the question text by means of the MetaMap Parser[16]. Developed by the National Library of Medicine, the MetaMap parser maps free text to concepts found within the UMLS Metathesaurus, in which each concept is uniquely identified by a given CUI. Thus, each LimeSurvey question is parsed by the MetaMap, returning a list of CUIs describing semantic concepts contained within the text. In order to query for CDEs, the returned list of CUIs has to be converted into NCI Thesaurus codes via the UMLS Terminology Services. Once the list has been successfully converted, the entire CDE database is queried for those CDE data elements who's "Object" or "Property" codes contain at least one of the extracted codes from the question text. The initial list of queried CDE data elements is then culled based on the number of total number of codes each discovered CDE shares with the question text, resulting in a final set of CDEs assured (by means of the MetaMap and NCI Thesaurus) to be the most semantically similar to the concepts in the question text.

Once all CDE data elements are discovered, and the syntax profile for the LimeSurvey question is initialized, the final step, the distance calculation, can be performed. In this step, the value domain of each discovered CDE data element is compared against the syntax profile of the LimeSurvey question being processed. The syntactic distance that exists between the LimeSurvey survey question being processed and a given CDE is defined by nine different distance calculations: minimum value distance, maximum value distance, precision distance, minimum character length distance, maximum character length distance, data type distance, enumerated value count distance, enumerated value text distance, and question text distance. Each respective distance calculation is shown below:

$$\boldsymbol{DistMinVal}(minVal_{CDE}, minVal_{LSQ}) = \boldsymbol{NormDis}(minVal_{CDE}, minVal_{LSQ}) \tag{1}$$

$$\boldsymbol{DistMaxVal}(maxVal_{CDE}, maxVal_{LSQ}) = \boldsymbol{NormDis}(maxVal_{CDE}, maxVal_{LSQ}) \tag{2}$$

$$\boldsymbol{DistPrec}(Prec_{CDE}, Prec_{LSQ}) = \boldsymbol{NormDis}(Prec_{CDE}, Prec_{LSQ}) \tag{3}$$

$$\boldsymbol{DistMinChar}(minChar_{CDE}, minChar_{LSQ}) = \boldsymbol{NormDis}(minChar_{CDE}, minChar_{LSQ}) \tag{4}$$

$$\boldsymbol{DistMaxChar}(maxChar_{CDE}, maxChar_{LSQ}) = \boldsymbol{NormDis}(maxChar_{CDE}, maxChar_{LSQ}) \tag{5}$$

$$\boldsymbol{DistDataType}(dataType_{CDE}, dataType_{LSQ}) = 1.0 - \|dataType_{CDE} \cap \boldsymbol{Map}(dataType_{LSQ})\| \tag{6}$$

$$\boldsymbol{DistEnumCount}(Enum_{CDE}, Enum_{LSQ}) = \boldsymbol{NormDis}(\|Enum_{CDE}\|, \|Enum_{LSQ}\|) \tag{7}$$

$$\boldsymbol{DistEnum}(Enum_{CDE}, Enum_{LSQ}) = \frac{\sum_{i=0}^{\|Enum_{LSQ}\|} \sum_{j=0}^{\|Enum_{CDE}\|} \boldsymbol{Min}\left(\boldsymbol{SmithWaterman}\left(Enum_{CDE_i}, Enum_{LSQ_j}\right)\right)}{\boldsymbol{Min}(\|Enum_{CDE}\|, \|Enum_{LSQ}\|)} \tag{8}$$

$$\boldsymbol{DistText}(Text_{CDE}, Text_{LSQ}) = \boldsymbol{Min}\left(\sum_{i=0}^{\|Text_{CDE}\|} \sum_{j=0}^{\|Text_{LSQ}\|} \boldsymbol{SmithWaterman}\left(Text_{CDE_i}, Text_{LSQ_j}\right)\right) \tag{9}$$

In the above equations, all values pertaining to a CDE value domain are denoted with the 'CDE' subscript; all values pertaining to a LimeSurvey survey question syntax profile are denoted with the 'LSQ' subscript. Functions **Min** and

Max refer to the functions that, respectively, return the minimum and maximum value from either two arguments or from a set of supplied numbers. **Map** refers to a function that accepts a LimeSurvey data type "hint" or a condensed data type (i.e. string, numeric, or enumerated) and returns the set of CDE data types mapped to the given LimeSurvey data type or condensed type. **SmithWaterman** refers to a generic implementation of the classical Smith-Waterman alignment algorithm[14] returning the alignment score divided by the maximum character length of the two supplied strings. **NormDis** is defined by the equation below:

$$\boldsymbol{NormDis}(a,b) = 1.0 - \frac{\mathbf{Min}(a,b)}{\mathbf{Max}(a,b)} \tag{10}$$

Granted the above, Equations 1-5 calculate the minimum value distance, maximum value distance, precision distance, minimum character distance, and maximum character distance by calculating the normalized distance (Equation 11) that exists between each value. If either the CDE value domain or the LimeSurvey question syntax profile does not contain values for one of aforementioned attributes, the maximum distance is assumed (normalized to 1.0) unless both the CDE value domain and the LimeSurvey question both lack a value for the same attribute. For instance, if both the syntax profile and the CDE value domain represent a string data type, both the syntax profile and the CDE value domain will be lacking all numerical constraints, such as maximum value. In this instance the minimum distance is assumed (normalized to 0.0).

Equation 6 calculates the distance that exists between the CDE value domain's data type and the data type defined by the LimeSurvey question's syntax profile. Using the **Map** function described in the prior paragraphs, the syntax profile's data type can be mapped to a defined set of CDE data types. With the generated set, the distance equation simply performs an intersection between the CDE value domain's data type and the generated set to determine if the syntax profile's data type is related to the CDE value domain's data type. The size of the resulting set (which will either be 1 if the two data types are related or 0 if they are disjoint) is subtracted by 1 to generate the normalized data type score.

Equation 7 calculates the normalized distance between the count of enumerated values of the CDE value domain and the count of enumerated values of the question syntax profile. Equation 8 calculates the normalized accumulated minimum text difference between each enumerated value from both the CDE value domain and the question syntax profile. To do this, the equation determines the minimum text alignment between two components from both the CDE value domain enumerated values list and the question syntax profile enumerated values list. This minimum alignment is summated with all other minimum alignments, with the resulting summation divided by the minimum of the size of the CDE value domain enumerated values list and the size of the question syntax profile enumerated values list. Equations 7 and 8 only apply to instances in which both or either CDE value domain and question syntax profile represent enumerated data. Just as with equations 1-5, if neither the value domain nor the syntax profile represents enumerated data, the minimum distance is assumed; if only one of either the value domain or the syntax profile represents enumerated data, the maximum distance is assumed.

Finally, Equation 9 simply finds the minimum alignment distance that exists between the LimeSurvey survey question's tokenized question text and the CDE value domain's name, question text, or preferred definition.

Once all syntactic distance calculations are performed for a given CDE data element and a LimeSurvey survey question, all resulting values are weighted and added together, resulting in a final normalized score produced between 0.0 and 1.0, with 0.0 representing a CDE data element-LimeSurvey question pair that is assumed to be syntactically and semantically identical. This 4-step process of culling CDEs semantically and calculating the syntactic distance is repeated for every CDE data element discovered with a given set of search terms, for each set of search terms generated from a given LimeSurvey question, for each LimeSurvey question from a given LimeSurvey survey. The CDE data element with the minimum score for a given LimeSurvey question is suggested to be the proper annotation for that LimeSurvey question, with each annotation detailing the discovered syntax and semantic profiles of the question as well as the value domain and semantic profiles of the discovered CDE.

CDE Importer

To apply syntactic constraints and annotate new clinical surveys with CDEs, the second tool, the CDE Importer, was developed as an plug in for the LimeSurvey to allow users to browse for and insert constraints defined by CDEs for each of their survey questions, explicitly annotating the survey questions with the CDE syntactic and, implicitly,

semantic information in the same process. The four primary steps for annotating new survey questions with the CDE Browser include 1: browsing for and selecting a CDE using search terms, 2: discovering any associated questions defined for the selected CDE, 3: browsing for and selecting from the list of returned associated questions (if there are any), and 4: review constraints and finalize any insertion options. Screenshots of the application executing each of these four tasks is shown in Figure 3 below:

To initiate the CDE Browser application, the user must first create a new question using the LimeSurvey survey software. The CDE Browser application modifies the LimeSurvey software such that it disallows a user to create a question without an associated CDE, thus requiring that all questions have proper CDE annotations. Once the user has started to create a new question, they can choose to browse for CDEs by activating an added button next to the LimeSurvey question "Code" field (the CDE Browser inserts the selected CDE's public ID as the code for a given question). This will generate a new window much like the one shown in frame 1 of Figure 3.

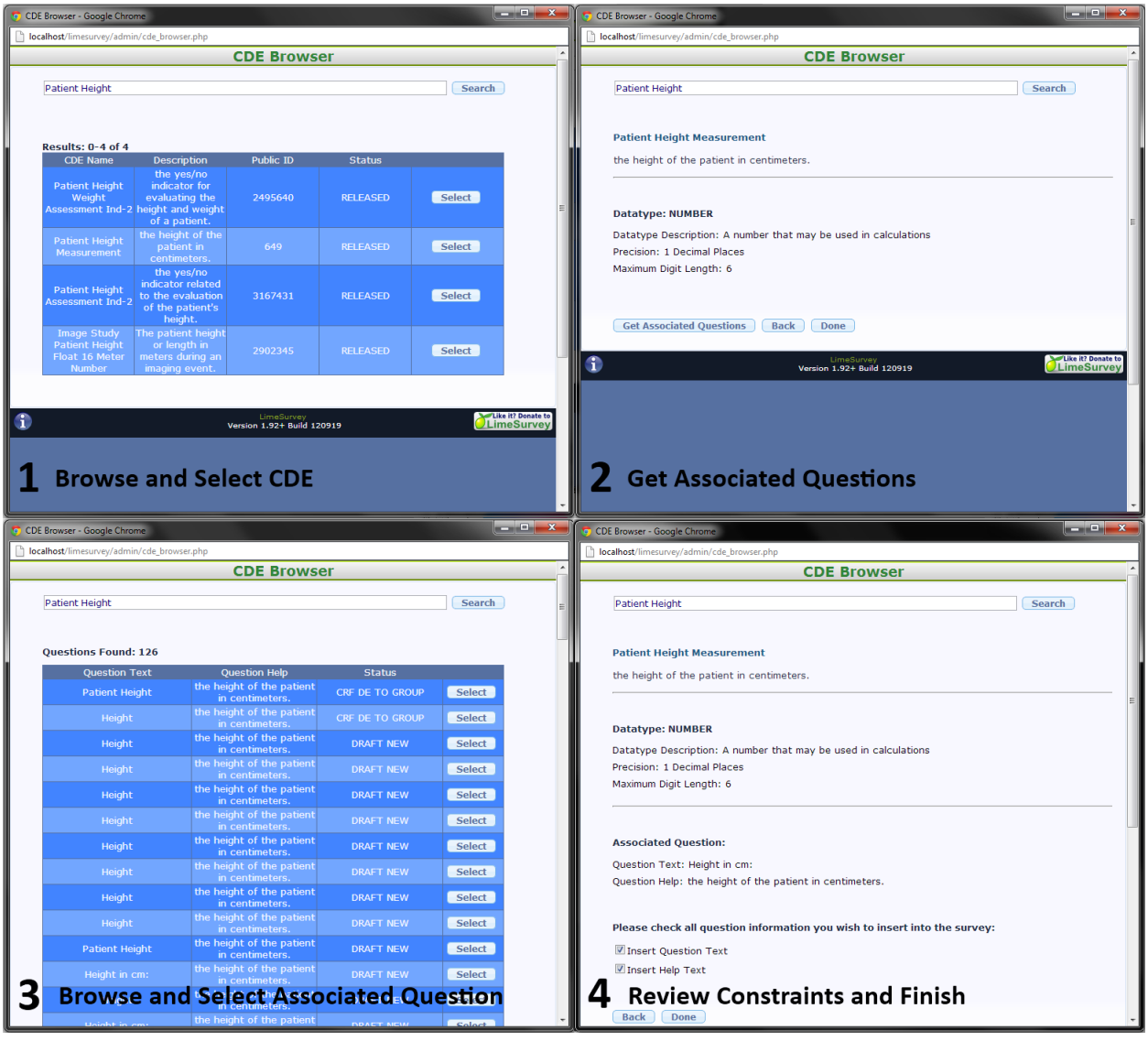


Figure 3. CDE Importer for LimeSurvey

Once the user has opened the CDE Browser window, they can then supply any number of search terms and hit "search." Utilizing the caDSR REST API provided by NCI, the CDE Browser will return the name, description, public ID, and status of each CDE, allowing the user to judge the appropriate CDE for their needs. If the user finds a CDE that is satisfactory, they can select it, bringing them to a window similar to frame 2 of Figure 3. The value

domain, or syntactic description, of the selected CDE is shown, describing the data type, a description of the data type, and any constraints imposed by the selected CDE (such as maximum or minimum precision). If the selected CDE has an enumerated value domain, all permissible values for that CDE are shown with the option to select which permissible answers the user wants to insert into the survey (as many enumerated value domains contain overlapping values and/or redundant values).

The user can choose to stop at step 2, or they can complete the CDE annotation process by activating the "Get Associated Questions" button. Pressing this will search for any questions that are related to the selected CDE using the caDSR Rest API. If any associated questions are found, a screen similar to frame 3 in Figure 3 is shown, displaying each question's preferred text, the question's help text, and the question's status. Just as with the CDE data element results, the user can use the displayed information to select the most appropriate question. Doing so will bring the user to a window much like frame 4, which again displays the CDE data element attributes and value domain, as well as the selected question text/help text with the options to insert these texts into the survey data. Selecting "Done" will automatically select the appropriate survey question type (based on the selected CDE's value domain), insert any enumerated values the user has chosen to insert, insert all syntactic constraints defined by the CDE's value domain, and insert any question text/help text the user has chosen to insert into the appropriate database table supplied by the LimeSurvey software. At this point, the survey question is considered to be annotated and thus the LimeSurvey software will allow the new question to be added to the current survey.

Results

To test the effectiveness of the Automated Annotation Tool, a random LimeSurvey survey extracted from data captured with UAMS's CRIS was analyzed with Automated Annotation Tool. The survey contained 80 questions and was manually annotated with CDEs to provide a ground truth for what is determined to be a "correct" annotation. Each manual designation was either given a 'weak' or 'strong' flag, describing, respectively, if the annotation almost exactly described the LimeSurvey question or if the annotation is only weakly semantically related. The only tool used to manually discover these CDEs was the aforementioned NCI's CDE Browser (utilizing simple text search). Of the 80 questions, 49 were determined to be weak annotations.

Once each Limesurvey survey question was given a "ground truth" annotation, the AAT was run on the survey, producing annotations that were determined to either be an exact match, semantically related, or a complete miss. An example of a random selection of results spanning these three types of annotations is shown in Table 1 below:

Table 1 Sample Automated Annotation Tool Results

Question Text	Common Data Element	Assessment
Meeting Code	Coding Scheme Identifier	Exact Match
What is your annual household income from all sources?	Patient Household Annual Income Amount	Exact Match
Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare?	Registration Private Health Insurance Medicare Payment	Semantically Related
When was the last time you had an EKG?	Number of months to Last Clinical Assessment	Semantically Related
At least once a week, do you engage in regular activity such as brisk walking, jogging, bicycling, or another activity long enough to work up a sweat?	Blackout Week Day Week	Complete Miss
You are afraid of finding colon cancer if you were checked. Would you say	Malignant Neoplasm Biopsy Finding Indicator	Complete Miss

you...		

In Table 1, one can see examples of the three types of annotations. The first and most desirable annotation type is naturally an exact match, in which the CDE annotation fully describes both the semantics and syntax of a given survey question. For example, "What is your annual household income from all sources," is both semantically and syntactically described by the CDE 'Patient Household Annual Income Amount.' In many cases, however, the AAT cannot determine an exact match but instead returns a CDE that is weakly semantically related to the question text. For instance, the CDE "Number of Months to Last Clinical Assessment" is weakly semantically related to "When was the last time you had an EKG?" as both refer to an elapsed time frame concerning a medical assessment, but the CDE does not specifically refer to an EKG assessment. Finally, a complete miss describes an annotation in which the annotation supplied describes neither the syntax nor the semantics of the survey question, two examples of which can be seen in Table 1.

Table 2 below shows the results of running the AAT on the entire survey, giving the percentage of the 80 annotated questions that were said to be annotated with an exact match, a semantically related match, or a complete miss. The first row shows results that include survey questions that could not be manually discovered (i.e. "weak" annotations). The second row shows the annotation results on only those questions that could be strongly manually annotated.

Table 2 Automated Annotation Tool Total Results

Adjusted	Exact Match	Semantically Related	Complete Miss
No	12.987%	14.285%	72.727%
Yes	28.571%	17.857%	53.571%

Conclusion

In this work, two tools, the Automated Annotation Tool and CDE Importer, are proposed and implemented to provide semantic and syntactic annotation for clinical research data to improve data quality of past clinical research data and constrain new clinical research to standard syntactic representations of survey questions and data elements. To test the effectiveness of the implemented Automated Annotation Tool, the tool was run against a randomly selected survey generated and maintained by UAMS's CRIS. In general, as the samples in Table 1 and the results in Table 2 expressed, a small portion of suggested annotations are syntactically and semantically sound, however many of the results are complete misses. It is the thoughts of this team that this is possibly due to disjoint semantic information between what is extracted by the MetaMap, the NCI Thesaurus, and UMLS. Another possibility is that certain semantic codes could overpower other codes. For instance, in Table 1, the question text mentions 'week,' forcing the AAT to focus on 'week' as a concept code, resulting in an incorrect annotation 'Blackout Day of the Week." Granted this, the next aim of this work is to provide a more powerful mechanic for determining which of the extracted codes from a given question text is to be considered more relevant given the context of the question in order to determine a more semantically related pool of potential CDEs. Another aim is to remove reliance on syntactic analysis for determining which of the semantically culled CDEs are the most "correct," as there is often very little correlation between the syntax profile of the survey question and the CDE syntax, even between correct annotations. Despite these shortcomings, the concepts exhibited by the tools implemented in this work display potential for future use and improvement for the goal of providing automated data quality assessment, improvement, and constraints for clinical research data.

Acknowledgements

This work was sponsored by the Winthrop P. Rockefeller Cancer Institute and by the award number UL1TR000039 from the National Center for Advancing Translational Sciences (NCATS). We also would like to thank University of Arkansas at Little Rock Information Quality Graduate Program for their support and guidance.

References

1. Gendron MS, D'Onofrio MJ. Data Quality in the Healthcare Industry. Data Quality [Internet]. 2001 Sep; 7(1) Available from: http://www.dataquality.com/901GD.htm
2. Health Level Seven International [Internet]. [publisher unknown]. [updated 2013; cited 2013 Mar 10]. Available from: http://www.hl7.org/

3. Covitx PA et al. caCORE: A common infrastructure for cancer informatics. Bioinformatics. 2003; 19(18):2404-12.
4. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004; 32(1):267-70.
5. Bodenreider O, Mitchell JA, McCray AT, Evaluation of the UMLS as a terminology and knowledge resource for biomedical informatics. AMIA. 2002; 61-65.
6. Richesson RL, Krischer J. Data Standards in Clinical Research: Gaps, Overlaps, Challenges and Future Directions. J AM Med Inform Assoc. 2007; 14: 687-96.
7. Mead CN. Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? J Health Inf Manag. 2006; 20(10): 71-8.
8. Tobias J et al. The CAP cancer protocols – a case study of caCORE based data standards implementation to integrate with the Cancer Biomedical Informatics Grid. BMC Medical Informatics and Decision Making. 2006; 6(25).
9. Wiley A. CTEP Common Data Elements [Internet]. [publisher unknown]. [updated 2012 Mar 16; cited 2013 Mar 10]. Available from: https://wiki.nci.nih.gov/display/caDSR/CTEP+Common+Data+Elements
10. McGilvray D. Executing Data Quality Projects: ten steps to quality data and trusted information. Massachusetts: Morgan Kaufmann Publishers; 2008. p. 19 – 23.
11. Patel AA 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).
12. Mohanty SK et al. The development and deployment of Common Data Elements for tissue banks for translational research in cancer – An emerging standard based approach for the Mesothelioma Virtual Tissue Bank. BMC Cancer [Internet]. 2008; 8(91). Available from http://www.biomedcentral.com/1471-2407/8/91.
13. Nadkarmi PM, Brandt CA. The Common Data Elements for Cancer Research: Remarks on Functions and Structure. Methods Inf Med. 2006; 45(6): 594-601.
14. Smith TF, Waterman MS. Identification of Common Molecular Subsequences. Journal of Molecular Biology. 1981; 147: 195-197.
15. Pushkarev V. Information Quality in Clinical Research Survey. UALR Information Quality Program Graduate Project. April 2010.
16. Aronson, A. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. AMIA 2001; 17-21.

An Author Topic Analysis of Tobacco Regulation Investigators

Ding Cheng Li, PhD[1], Janet Okamoto, PhD[2], Scott Leischow, PhD[2], Hongfang Liu, PhD[1]
Mayo Clinic, Rochester, MN[1], Mayo Clinic, Phoenix, Arizona[2]

Abstract. *To facilitate the implementation of the Family Smoking Prevention and Tobacco Control Act of 2009, the Federal Drug Agency (FDA) Center for Tobacco Products (CTP) has identified research priorities under the umbrella of tobacco regulatory science (TRS). As a newly introduced field, the current landscape of TRS research is unclear. In this work, we conducted a bibliometric study of TRS research by applying author topic modeling on MEDLINE citations published by currently-funded TRS principle investigators. Our initial results show that author topic modeling can address the issue of research interests reasonably. Furthermore, a network involving authors, topics and words can be established for more detailed bibliometric analysis. This network may also be useful to grantees and funding administrators in suggesting potential collaborators or identifying those that share common research interests for data harmonization or other purposes.*

Background and Introduction. To facilitate the implementation of the Family Smoking Prevention and Tobacco Control Act of 2009, the Federal Drug Agency (FDA) Center for Tobacco Products (CTP) was formed to oversee tobacco regulatory activities. Its responsibilities include setting performance standards, reviewing premarket applications for new and modified risk tobacco products, requiring new warning labels, and establishing and enforcing advertising and promotion restrictions. In order to meet these responsibilities, the CTP has identified research priories for tobacco regulatory science (TRS) in order to inform and guide the CTP's regulatory decision-making. While tobacco researchers have been examining some of the CTP's TRS research priorities for many years, they have not necessarily been doing so under the umbrella or specific title of 'tobacco regulatory science'. Therefore, examining and identifying research topics from the corpus of TRS work could help to more clearly define this growing research area. In this paper, we applied author topic modeling [1], a variation of Latent Dirichlet Allocation (LDA), to simultaneously model the content of documents and the interests of authors. Namely, given the broader TRS research field, we attempted to discover topics as well as general research interests utilizing MEDLINE citations for currently funded TRS investigators.

Methods. We obtained all MedLINE citations published by the principle investigators (PIs, 133 in total) of TRS grants funded by the CTP through Tobacco Regulatory Science Research Program (TRSP) (http://prevention.nih.gov/tobacco/portfolio.aspx). Since each article can have multiple authors, the author set considered here are PIs (can appear in any place in the paper) plus the last author of the paper. The final author set includes 2,740 authors. The document set includes those MEDLINE citations with abstract available, resulting in 7460 abstracts. For each document, we remove stop words using a stop word list available at Mallet software package. We further filter words based on Term Frequency-Inverse Document Frequency (TF-IDF), where words with high document frequencies and relatively insignificant for single document are removed. We then stem the words by applying the potter stemmer. We ran the author topic modeling developed by (11) on it for 50 iterations. Topic number T is selected as 20. The hyperparameters α and β are fixed as 50/T and 0.01 respectively.

Preliminary Results. The results, which yield 20 topics in Figure 1 show that this approach can efficiently cluster collections of articles into discriminative categories without any supervision. More interestingly, it can associate topics to authors in a high accuracy. This indicates that we may incorporate author topic modeling into author identification systems to infer the identity of an author of articles using topics generated by the model. The relevance of this analysis to TRS is at least twofold. First, this analysis is a 'proof of concept' that can be beneficial assess the change over time in TRS as new projects are funded and collaborative science in this area changes. The results can thus we used to assess the extent to which new research reflects the funding priorities of the FDA. Second, author topic modeling outcomes can be used by investigators to assess who is conducting research in a particular research domain in order to foster collaborative science. By fostering collaborative science in TRS, it becomes possible to speed advances in that science by fostering communication between scientists that can avoiding un-needed duplication and impact decision-making on new science that can benefit regulatory decision-making.

Reference:

[1] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 2004, pp. 487-494.

[2] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, and X. Zhu, "Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010," *Journal of the American Medical Informatics Association,* vol. 18, pp. 557-562, 2011.

Figure 1 Word cloud for top words of 20 topics

Developing Governance for Federated Community-based EHR Data Sharing

Ching-Ping Lin PhD[1,2], Kari A. Stephens PhD[1], Laura-Mae Baldwin MD MPH[1], Gina A. Keppel MPH[1], Ron J. Whitener JD[1], Abigail Echo-Hawk MPS[1], Diane Korngiebel DPhil[1]
[1]Institute of Translational Health Sciences, University of Washington, Seattle, WA,
[2]Global REACH, University of Michigan, Ann Arbor, MI

Abstract

Bi-directional translational pathways between scientific discoveries and primary care are crucial for improving individual patient care and population health. The Data QUEST pilot project is a program supporting data sharing amongst community based primary care practices and is built on a technical infrastructure to share electronic health record data. We developed a set of governance requirements from interviewing and collaborating with partner organizations. Recommendations from our partner organizations included: 1) partner organizations can physically terminate the link to the data sharing network and only approved data exits the local site; 2) partner organizations must approve or reject each query; 3) partner organizations and researchers must respect local processes, resource restrictions, and infrastructures; and 4) partner organizations can be seamlessly added and removed from any individual data sharing query or the entire network.

Introduction

A key aim of the National Institutes of Health (NIH) Roadmap has been to broaden the participation of communities and practice-based care settings in medical and health services.[1] Bi-directional translational pathways between scientific discoveries and primary care are crucial for improving individual patient care and population health.[2]

The Institute of Translational Health Sciences (ITHS) developed the Data QUEST (Data QUery Extraction Standardization and Translation) pilot project to build a technical infrastructure to support the sharing of electronic health record (EHR) data across primary care practices and tribal communities.[3, 4] Data QUEST targets engagement with disparate primary care practice based organizations serving small rural populations because they are often excluded from research and dissemination efforts.[5]

We iterated our technical design with our community-based partners, five organizations in the WWAMI region Practice and Research Network (WPRN) and five American Indian/Alaska Native (AI/AN) tribal practices, the ITHS-based community liaisons, and project investigators,.[6, 7] While gathering technical requirements, we also deliberately engaged in a time and resource intensive process to learn about governance requirements. We knew that the success of our project depended on our ability to foster trust and consider complicated and independent governance issues across community based practice partners, many of which served underserved populations and some of which involved communities with tribal sovereignty.

Background

Data governance is defined as the process by which the responsibilities of data stewardship (the acquisition, storage, aggregation, de-identification, release, and use of data) are conceptualized and carried out through policies and approaches.[8] Data governance is crucial for maintaining privacy protection for community members at the individual and group level. The research community may be familiar with HIPAA protections that address individual data privacy through protected health information, but anonymous aggregated data across groups can also be damaging by characterizing subsets of patient populations leading to stigmatization (e.g., substance abuse, sexually transmitted infection, mental illness, obesity rates) and thus need similar governance policies.

Creating new data sharing pathways must consider thoughtful changes to data governance, which involve high levels of investment and trust by partner organizations. Data sharing systems can enforce different governance requirements through authentication models and levels of automation for querying and receiving data.[9-12] However, governance requirements are independent of any specific implementation approach. We needed to establish proper governance requirements within Data QUEST that met the needs of partner organizations before determining our implementation approach.

Methods

We reviewed requirements for gathering data with five WPRN partner organizations and five American Indian/Alaska Native (AI/AN) tribal practices located across the WWAMI (Washington (n = 5), Wyoming (n = 0) Alaska (n = 2), Montana (n = 1), Idaho (n = 2)) region. The purpose of these discussions was to determine both technical and governance readiness for Data QUEST. On average, each non-tribal practice (primary care community-based practice, n = 5) supported 30-35 clinical providers per organization, variably dividing their time between clinical, administrative, and teaching duties. The AI/AN partners (n = 5) included two general types of practices: practices for which tribes received funding from the United States government for clinical operations, but were managed by the local tribal administration and governance and practices funded and managed by the Indian Health Service, a division of the United States Health and Human Services. They supported a range of 3-5 providers generally working full time and served a similar volume of visits per year (ranged from 10,000-40,000).

We spoke to a diverse set of practice leaders, including clinicians, technical staff, and administrators to identify stakeholders and leaders necessary to support and authorize a data sharing project. We presented current governance practices based on existing practice based research networks and their related data sharing efforts and tools to drive discussions of governance requirements (i.e., Group Health Research Institute's Virtual Data Warehouse[10], DARTNet Institute's data sharing infrastructure[12]). While working with tribal based practices, we also consulted with tribal leaders and received approval by the tribal governing bodies. We recorded feedback from community liaison and investigator partners.[7] As we cycled through iterations with partners in developing the Data QUEST technical architecture, we addressed technical implications of the governance requirements.[13] We recorded local governance and engagement procedures, as well as engaged partners in iterating governance requirements for Data QUEST data sharing. Qualitative data gathered through this iterative design process defined governance requirements.

Results

From continued engagement with partners, we identified four governance requirements and governing principles that we outline below:

Governance Requirement 1: Organizations can physically terminate the link to the network and only approved data leaves the local site.

Two common approaches to data sharing are: 1) a *centralized* approach in which data from each organization are aggregated and stored in a single physical repository for inquiry and 2) a *federated* approach in which the data remain at each practice with access to the data through a virtual repository.[14] Our partners clearly preferred a federated approach.

Interviews with partners revealed concerns over data ownership, control, and security of identifiable data. In centralized architectures, while partners can approve or deny data requests either through technical or social means, they typically do not have physical control over the data and servers, which are managed by the centralized data steward. In a federated system, if necessary, the local data manager can simply physically turn off the database or server, and the local data can no longer be shared. This capability to "flip the off switch" was important to our partners, given that the identified data included Protected Health Information. Another motivation for a federated approach was the desire to only release sanctioned data to the outside as opposed to releasing all the possible sharable data into a centralized repository and rely on a centralized data steward to facilitate access.

Figure 1 shows Data QUEST's federated architecture. Organization A, Organization B, and Organization C are part of the Data QUEST network, and all the data reside locally at the organization (within the boundaries of the local firewall system) until they decide to share data with outside partners for a specific query or project. No centralized warehouse or database contains all of the data in the network. The "combined health data" are aggregated on a project-by-project basis.

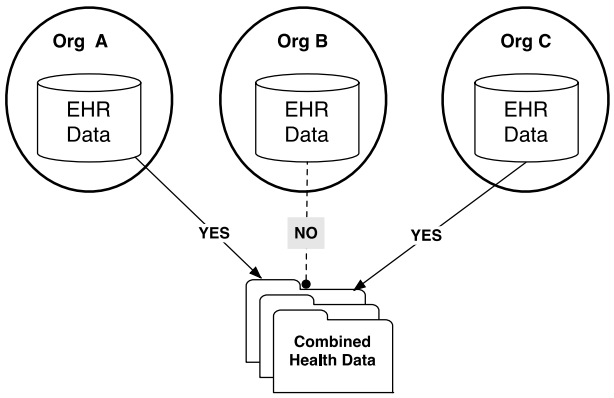


Figure 1. Governance Requirement 1 (physically terminate the link to the network and only approved data leaves the firewall), and Governance Requirement 2 (approve or reject each query)

Governance Requirement 2: Partners must approve or reject each query

Our partners expressed concern that users of the Data QUEST network might "data fish," that is, download and analyze health data without a constructive or approved research question. Data fishing could cause exploitation or vulnerabilities, due to sensitive disease-based issues that may stigmatize communities or reveal sensitive clinical quality issues. Tribal communities related these preferences particularly in the context of historical contentious relationships with outside researchers, especially in the area of data sharing and publication of results without community oversight, and all of our partners reflected these preferences.[15] Data fishing raises many ethical issues that can be addressed by federation and careful control over data ownership. Our partners stated ownership of the data, identified or unidentified, must remain in the hands of the community or practice.

Our partners also stated they must be able to review every data query request and result, with the option to deny access and/or withdraw from participation at any time, in addition to Institutional Review Board (IRB) or prior approval. Therefore, all data query requests must pass through a designated authority at the partner organization who must explicitly approve all queries before results are delivered.

Governance Requirement 3: Partner organizations and researchers must respect local processes, resource restrictions, and infrastructures

Each practice and community has unique approval processes, human resources, and different methods for engagement. From gathering information on local processes and social structures, we recognized our third governance requirement to respect local processes, resource restrictions, and infrastructures. This requirement is reflected through a set of core data sharing documents that govern data sharing:

- Data Use Agreement (DUA) authorizing the sharing of the health data with partners
- Publication Policy (PP) outlining how publications and presentations presenting data will be vetted
- Memorandum of Understanding (MOU) creating an agreement to participate in specific research projects

In addition to DUAs, PPs, and MOUs, each organization also required Business Associate Agreements with our vendor, Data Transfer Agreements between the vendor and the organization, and contracts between the university and the vendor.

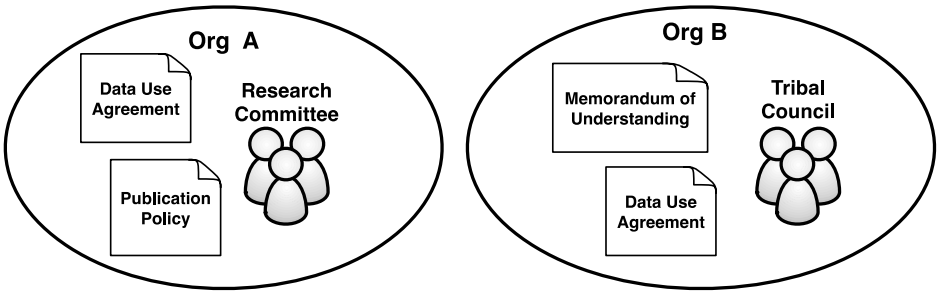


Figure 2. Heterogeneous local governance infrastructure examples

Figure 2 shows an example of two Data QUEST organizations that have different data sharing governance structures in terms of documents, policies, and governing boards. At both organizations A and B, we have developed Data Use Agreements. Organization B required a Memorandum of Understanding before we could begin to implement our data sharing infrastructure. Our tribal partners required an additional publication policy that the other partner organizations subsequently adapted and adopted. These policies detailed how publications, presentations and abstracts will be vetted, how communities will be described in manuscripts, and an arbitration system for disputes. Figure 2 shows that Organization A uses a research regulatory committee to oversee research activities, while at Organization B, the Tribal Council oversees research activities related to the tribal clinic. The implementation teams for the AI/AN communities and WPRN worked together to share governance materials and processes that facilitated development. Commonalities were identified and eventually we produced specific documents to be used by each group (e.g., DUAs and PPs).

Developing Data QUEST required working within these local processes. Any researchers wishing to partner with Data QUEST must also respect and comply with these established processes.

Governance Requirement 4: Organizations can be seamlessly added and removed from data queries or from Data QUEST itself

The goal of the Data QUEST pilot project was to start with a small initial number of community-based partners to develop the foundational infrastructure and proof of concept for a data sharing network, with the intention to grow. We have also engaged national efforts, most notably with the WPRN joining the DARTNet Institute,[12] expanding our data sharing capacity nationally. Anticipating the potential need and desire to grow partnerships and engage in parallel / national efforts, we recognized that the governance process must be flexible enough to support the ability to add (or remove) Data QUEST organizations without interrupting data sharing capabilities for our participating organizations.

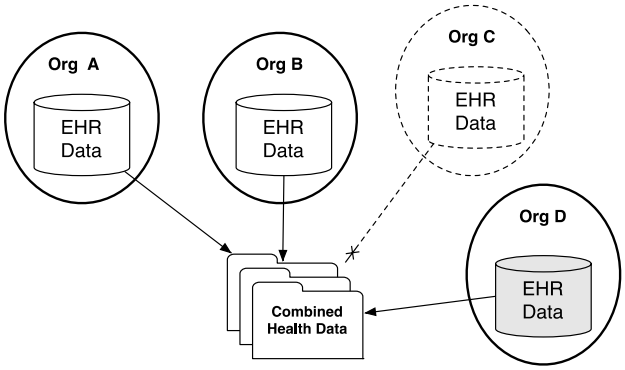


Figure 3. Data QUEST's governance must permit the addition (Org D) and subtraction (Org C) of organizations easily from individual data sharing projects or from Data QUEST itself.

Partner organizations do not have the ability to vote additional partners in or out, nor have any influence over other organizations' participation in Data QUEST or in any individual query or project. Instead, all partners are allowed to participate or withdraw at any time for any individual data query / project or from Data QUEST. This flexibility is easily facilitated because of the network's federated architecture as discussed above. In Figure 3, Org D has decided to join the network and Org C has decided to withdraw from the network. Org C is able to withdraw by simply disconnecting. This "disconnection" may be a physical technical connection that can be turned "off" or it may be an abstract connection with the network or both.

Discussion

Community-based partners require local control over their data and are reluctant to allow their data to reside permanently outside their domain of control, even if they have full oversight over the data. In addition, building trust between academic research and community-based partners can be facilitated by developing systems that maximize local control and allowing for expectations of high granular control over the flow of data to researchers. Community-based partners require the ability to control the data flow at the individual query level, not just at the research project level, which may include several queries, though they may choose to grant access to data on a per project basis. Because the Data QUEST network consists of disparate and geographically distant organizations without shared governance, some of which involve tribal nations with complex legal requirements, we needed to

understand and support each partner's local governance infrastructure and research processes, including our own institution's requirements for facilitating research or receiving health data from non-affiliated institutions. This process was time consuming and critical to the success of our network development efforts.

Multiple approaches can fulfill governance requirements

It is important to re-iterate that none of these requirements define a specific *technical* solution. They define a set of parameters that any automated or non-automated solution must meet, but they leave open the possibility for multiple approaches. For instance, Governance Requirement 2 (partners must approve or reject every query) can be met through a software functionality where the organization can view and approve the query electronically or through a manual, people-based process.

We did not develop customized data sharing software for Data QUEST due to resource restraints. However, other efforts are underway to develop and distribute software that instantiates many of these requirements (e.g. i2b2[11], PopMedNet[16]). The complexity of adapting one of these budding technologies to meet our immediate governance requirements was not feasible. We therefore used a mixed social and technical approach to support the workflow and governance of the network.

Continuous and iterative engagement of partners is crucial

Determining governance and technical dimensions in primary care community-based settings for growing data sharing across EHR data is best addressed through iterative and inclusive involvement with community partners. Community partners must be engaged from the onset in developing technical requirements to ensure that governance requirements are incorporated and implementation of pre-built software meets their needs.

However, finer control also adds to the workload of each partner to manage their local database, queries from researchers, and proposed research projects potentially slowing down the functionality of the entire network while some sites wait on others for decisions. Additionally, Governance Requirements 2 and 3 also create a direct relationship between the number of sites in the network and the efficiency of the system - the more sites, the more site-specific complexity is added to the network.

It takes significant time to meet with partners about their concerns and needs. However, these discussions are crucial because resources vary across organizations. For instance, one organization had a research project coordinator who could serve as a dedicated liaison to manage data requests, whereas another organization did not. This iterative process helped determine whether partners felt they could participate and revealed additional system requirements to be built in the context of each organization's existing resources. As data sharing networks and the volume of research partnerships grow, so does the need for managing data requests and communication between communities, practices, investigators, and information technology specialists. All stakeholders must be engaged in this process and iterate together.

Conclusions

We have presented the governance requirements for Data QUEST, a pilot project building a data sharing architecture of community-based and tribal primary care practices across the northwest region: 1) organizations can physically terminate the link to the network and only approved data leaves the firewall; 2) partners must approve or reject each query; 3) partner organizations and researchers must respect local processes, resource restrictions, and infrastructures; and 4) organizations can be seamlessly added and removed from a query or from the network itself. Support for these requirements may be automated or manual.

It is crucial to build data sharing networks in community-based settings so that translational science can succeed at bridging scientific discovery to front line treatment environments in primary care settings. Researchers can capitalize on these networks to speed significant health impacts by conducting work in these real-world settings. These networks can promote inclusion of underrepresented and rurally located people in research, who are so often missed in research conducted in academic-based settings. Dissemination science is growing and the need to conduct comparative effectiveness trials will be served well by the efficiencies offered within these complex data sharing environments. Time and resources are precious in these settings however, and care and iteration must be engaged to develop these networks effectively.

Acknowledgements

This publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR000423. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Many thanks to our partner organizations in the Data QUEST pilot program, Alfred O. Berg, Peter Tarczy-Hornoch, and our colleagues in the Community Outreach and Biomedical Informatics programs of the ITHS.

References

1. Zerhouni EA. US Biomedical Research: Basic, Translational, and Clinical Sciences. JAMA. 2005 September 21, 2005;**294**(11):1352-8.
2. Westfall JM, Mold J, Fagnan L. Practice-Based Research--"Blue Highways" on the NIH Roadmap. JAMA. 2007 January 24, 2007;**297**(4):403-6.
3. CTSA Principal Investigators. Preparedness of the CTSA's Structural and Scientific Assets to Support the Mission of the National Center for Advancing Translational Sciences (NCATS). Clinical and Translational Science.**5**(2):121-9.
4. Lin CP, Black RA, Laplante J, et al. Facilitating health data sharing across diverse practices and communities. AMIA Summits Transl Sci Proc. 2010;**2010**:16-20.
5. Hodge FS, Weinmann S, Roubideaux Y. Recruitment of American Indians and Alaska Natives into clinical trials. Ann Epidemiol. 2000 Nov;**10**(8 Suppl):S41-8.
6. Lin CP, Stephens KA, Black RA, et al. Facilitating Health Data Sharing Across Diverse Practices and Communities. 2010 AMIA Summit on Clinical Research Informatics. San Francisco, CA: AMIA; 2010.
7. Stephens KA, Anderson N, Lin CP. Developing best practices for evaluating federated data sharing: Approaches from academic hospital and primary care clinic networks. AMIA Annu Symp Proc. 2010.
8. Rosenbaum S. Data Governance and Stewardship: Designing Data Stewardship Entities and Advancing Data Access. Health Services Research.**45**(5p2):1442-55.
9. Lin K, Daemer G. caBIG Security Technology Evaluation White Paper: National CAncier Institute; 2006 January 23, 2006.
10. Vogt TM, Elston-Lafata J, Tolsma D, Greene SM. The role of research in integrated healthcare systems: the HMO Research Network. Am J Manag Care. 2004 Sep;**10**(9):643-8.
11. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). Journal of the American Medical Informatics Association. March 1, 2010;**17**(2):124-30.
12. Pace WD, Cifuentes M, Valuck RJ, et al. An electronic practice-based network for observational comparative effectiveness research. Ann Intern Med. 2009 Sep 1;**151**(5):338-40.
13. Stephens KA, Lin CP, Baldwin LM, et al. LC Data QUEST: A Technical Architecture for Community Federated Clinical Data Sharing. 2012 AMIA Summit on Clinical Research Informatics. San Francisco, CA: AMIA; 2012.
14. Schatz B, Mischo WH, Cole TW, et al. Federating Diverse Collections of Scientific Literature. Computer. 1996;**29**(5):28-36.
15. Whitener RJ. Research in Native American Communities in the Genetics Age: Can the Federal Data Sharing Statute of General Applicability and Tribal Control of Research Be Reconciled? J Tech L & Pol'y. 2010;**15**:217.
16. Brown JS, Holmes JH, Shah K, et al. Distributed health data networks: a practical and preferred approach to multi-institutional evaluations of comparative effectiveness, safety, and quality of care. Med Care. 2010 Jun;**48**(6 Suppl):S45-51.

Facilitating post-surgical complication detection through sublanguage analysis

Hongfang Liu, PhD[1], Sunghwan Sohn, PhD[1], Sean Murphy[1], Jenna Lovely, PharmD R.Ph[2], Matthew Burton, MD[1,2], James Naessens, ScD[1], David W Larson, MD[2]
[1]Department of Health Sciences Research [2] Department of Surgery
Mayo Clinic College of Medicine, Rochester, MN 55905

Abstract

Identification of postsurgical complications is the first step towards improving patient safety and health care quality as well as reducing heath care cost. Existing NLP-based approaches for retrieving postsurgical complications are based on search strategies. Here, we conduct a sublanguage analysis study using free text reports available for a cohort of patients with postsurgical complications identified manually to compare the keywords identified by subject matter experts with words/phrases automatically identified by sublanguage analysis. The results suggest that search-based approaches may miss some cases and the sublanguage analysis results can be used as a base to develop an information extraction system or support search-based NLP approaches by augmenting search queries.

Introduction

Identification of postsurgical complications is the first step towards improving patient safety and health care quality as well as reducing health care cost (1). Multiple methods currently are being used to identify patients with postsurgical complications. The Agency for Healthcare Research and Quality (AHRQ) Patient Safety Indicators (PSI) are based on administrative hospital ICD-9 discharge diagnoses enhanced with present on admission (POA) modifiers (2). The American College of Surgeons National Surgical Quality Improvement Program (NSQIP) bases its assessment on clinical registry data abstracted directly from a sample of medical records (3). Consumer Reports used POA-enhanced hospital claims data to identify cases with prolonged (i.e., longer than expected) risk-adjusted post-operative lengths of stay (prLOS), which served as a surrogate indicator for serious non-fatal inpatient adverse outcomes (4).

With the rapid adoption of electronic medical records (EMRs) and the accelerated advance of health information technology (HIT), detection of postsurgical complications based on EMRs via natural language processing (NLP) offers a potential powerful alternative to either administrative data or labor-intensive manual chart reviews (5, 6). For example, Murff et al. (6) showed NLP-based postsurgical complication detection achieves higher sensitivity and lower specificity compared to PSI when using a randomly selected sample of Veterans Affairs Surgical Quality Improvement Program (VASQIP)–reviewed surgical inpatient admissions. The NLP approach applied in their study is a generic SNOMED coding system followed by an inclusion-exclusion search. Alsara et al (7) applied a similar search strategy to identify pertinent risk factors for postsurgical acute lung injury.

Meanwhile, in the field of NLP, sublanguage-oriented information extraction (IE) has shown to be promising in extracting a variety of structured data from clinical reports (8, 9). In this paper, we conducted sublanguage analysis related to postsurgical complications utilizing clinical narratives related to a cohort of colorectal surgical patients to compare words/phrases identified by sublanguage analysis with a list of keywords identified by subject matter experts and to explore the potential of developing IE applications for postsurgical complication detection.

Background

Colorectal postsurgical complications

Postsurgical complications may be general or specific to the type of surgery undertaken. In this paper we studied complications that may happen after colorectal surgery. Table 1 shows the complications and their definition considered in this paper adapted from: American Society of Colon & Rectal Surgeons (http://www.fascrs.org/physicians/education/core_subjects/2011/Complications/) and Society of international radiology (http://www.sirweb.org/).

Table 1. Colorectal postsurgical complication definitions

Postsurgical Complication	Abbreviation	Definition
Deep vein thrombosis (DVT)/pulmonary embolism (PE)	DVTPE	DVT is the formation of a blood clot, known as thrombus that usually occurs in the leg although it can happen in other parts of the body. Part of the clot can break off and travel to the lung, where it blocks the oxygen supply, casing heart failure, known as PE.
Bleeding	BLEED	Anastomotic bleeding is common and varies in severity. More serious bleeding can be managed with epinephrine and saline retention enemas. If this fails, surgical intervention can be performed.
Wound infections	INFECTION	Wound infections occur in 5-15% of patients following colorectal surgery and typically present around the fifth postoperative day and treated by opening of the overlying skin incision.
Myocardial infraction	MI	Acute myocardial infarction occurring during surgery or within 30 days after surgery.
Ileus	ILEUS	Ileus is simply defined as bowel obstruction. CT scan of the abdomen and pelvis has the sensitivity of 90-100% for diagnosis and evaluation of small bowel obstruction.
Abscess/Leak	ABSCESS	Extravasation of contrast material limited to the perianastomotic space often results in the development of an abscess, a pocket of infected fluid and pus. This is usually managed by insertion of a radiologically-guided percutaneous drainage catheter. Anastomotic leak varies depending on the level of anastomosis. Small bowel and ileocolic anastomoses have the lowest rates and coloanal anastomoses have the highest rates.

Sublanguage analysis

The hypothesis behind an IE system is the property of inequalities of likelihood in the sublanguage. We focus on two kinds of information: domain taxonomy and semantic lexicon where domain taxonomy here refers to report types while semantic lexicon refers to a collection of words/phrases. Specifically, we argue that clinical text for patients with surgical complications can have different distribution regarding report types. In addition, words and phrases for patients with surgical complications can also have different distributions compared to those with no complications. Identifying words/phrases with high inequality of likelihood can be used to assist the knowledge engineering process of developing an IE system or formulating the queries to identify positives (here, postsurgical complications).

Materials and Experimental Methods

Colorectal surgical cohort

The cohort considered here contains 1,856 colorectal surgical cases for 1,416 patients between 2005 and 2013 enrolled at Mayo Clinic Rochester. For this cohort, a quality improvement project has documented the postsurgical complications defined in Table 1. The cohort has been used as a retrospective data set for developing an IE system for identifying postsurgical complications. A collection of keyword patterns relevant to specific postsurgical complications has been assembled in the period of six months by subject matter experts. Table 2 lists the current keyword patterns. Since there can be multiple surgical cases per patient, to reduce patient-specific sublanguage, we limit to one surgical case per patient in this study. For each patient, we chose the case with the latest surgical date in case of multiple surgical cases for that patient.

MedTagger

MedTagger is a concept mention detection and normalization tool released open source through open health natural language processing (10, 11). It consists of three components: dictionary lookup allowing flexible mapping, machine learning-based concept mention detection, and pattern-based information extraction. In this study, we used MedTagger to identify phrases present in MedLex, a general semantic lexicon created for the clinical domain (12).

Table 2. Keyword patterns (as regular expressions) identified by subject matter experts. "\W" means punctuations, "\w+" means one or more letters, "\s+" means one or more blank spaces, and P? means the pattern P occurs zero or once.

Postsurgical Complication	Keyword patterns
Deep vein thrombosis (DVT)/pulmonary embolism (PE)	dvt; vein(\W|\s+)?thrombosis; venous(\W|\s+)?thrombosis; venous(\W|\s+)?thromboembolism; vte; vena(\W|\s+)?cava thrombosis; pe; pulmonary(\W|\s+)?embol(\w+)?; pulmonary embol(\w+)?; pulmonary(\W|\s+)?thromboembol(\w+)?
Bleeding	bleed(\w+)?; hemorrhage; acute blood loss; acute(\W|\s+)?anemia; acute blood loss anemia; post.?op(\w+)? anemia
Wound infections	wound(\W|\s+)?infection; cellulitis; contamination within the abdomen
Myocardial infraction	ami; attacks? coronary; attacks? heart; cardiac infarctions?; coronary attacks?; heart attacked; heart attacks?; heart infarctions?; infarctions? myocardial; infarctions? of heart; infarctions? heart; infarcts? myocardial; myocardial(\W|\s+)?infarcts?; myocardial(\W|\s+)?infarctions?; myocardial necrosis
Ileus	ileus; enteric tube; naso(\W|\s+)?gastric; naso(\W|\s+)?enteric; ng; ngt; ng(\W|\s+)?tube; small(\W|\s+)?bowel obstruction; sbo; partial small(\W|\s+)?bowel obstruction; psbo; poi
Abscess/Leak	Abscess; intra(\W|\s+)?abdominal infection; intra(\W|\s+)?abd infection; abdominal infection; leak; anastomotic leak; fistul(\w+)?

Data processing and analysis

We extracted various free text reports within 30 days from the surgical dates for the cohort. We lowercased all words and acquired words mentioned in the reports. We then applied MedTagger to obtain clinical concept phrases mentioned in the reports. We considered words and phrases as candidate concept keywords.

For each complication, we applied the following equations adapted from point-wise mutual information (http://en.wikipedia.org/wiki/Mutual_information) to assess the inequality of likelihood of report types and words/phrases:

$$Inequality(rpt, com) = log2((N(rpt, com) + 0.01)/N(com)) - log2(N(rpt)/N) \quad \text{(Eq 1)}$$

$$Inequality(con, com) = log2\big(N(con, com)\big) * \left(log2\left(\frac{N(con,com)+0.01}{N(com)}\right) - log2\left(\frac{N(con)}{N}\right)\right) \quad \text{(Eq 2)}$$

where N is the number of surgical cases, $N\ (rpt)$ is the number of cases having report type rpt, $N\ (con)$ is the number of cases having concept con, $N(rpt, com)$ is the number of surgical cases with complication com and report type rpt, $N(con, com)$ is the number of surgical cases with complication com and concept con, and $N(com)$ be the number of surgical cases with complication com. Note that in Eq 2, we penalized those concepts with low co-occurrence with the complications.

Results and discussion

We retrieved a total of 23,558 reports with an average of 16.6 reports per patient. After excluding report types with lower than 100 occurrences, we computed the inequality measures of the remaining report types. Table 3 shows the statistics of reports and postsurgical complications. Figure 1 shows the inequality measures of report types where close to 0 indicates no difference of the distribution of the cases with or without the specific complication (computed using Eq 1). It indicates that surgical cases with postsurgical complication generally yield more reports and for certain report types such as radiology and ECG, we observe over two fold increase.

For words/phrases, we filtered out those occurred in less than three patients and obtained a total of 21,910 unique words/phrases. Table 4 lists the top words/phrases ranked according to inequality measures computed using Eq 2). When comparing Table 2 and Table 4, some of the patterns identified by subject matter experts are also ranked top

by sublanguage analysis. We underlined words/phrases captured by keyword patterns and also crafted by subject matter experts. Majority of the patterns identified in Table 2 do not appear in Table 4 and verse versa. One extreme case is INFECTION where none of the top ranked words/phrases can be found through keyword patterns. Another example is, *abscess*, a keyword for ABSCESS which appears in reports of 336 cases and only 20 are cases (with a rank of 345 for ABSCESS). Meanwhile, sublanguage analysis identified some keywords which can be a strong signal of complications. For example, *low hemoglobin* (32 out of 93 patients containing *low hemoglobin* in reports are cases) which ranked the third for BLEED most likely indicates bleeding. At the same time, the term, *bleeding,* appears in the reports of 408 patients but only 57 of them are bleeding cases (ranked 156 according to inequality). However, some of the keywords identified by sublanguage analysis may not have obvious semantic relationships with the postsurgical complications. For example, *knee* appears in five out of the seven cases of DVTPE with a total occurrence of 61. It ranks 13 for DVTPE regarding its inequality. But there is no obvious semantic relationship between *knee* and *DVTPE*. Further investigation is needed to find out hidden semantic relationships.

Table 3. Statistics of reports and postsurgical complications.

Report Type	Definition	#Patients	ABSCESS	MI	BLEED	AFIB	ILEUS	DVTPE	INFECTION
PRG	Progress notes	877	19	1	54	11	99	5	28
MIS	Misc. notes	887	20	3	56	11	87	7	38
SUM	Discharge summary	1404	23	5	78	17	136	7	46
OPN	Operation notes	1411	23	5	80	17	137	7	47
THP	Therapy	737	18	3	49	12	85	6	37
CON	Consultant notes	556	19	4	50	13	66	7	38
ADM	Admission	1003	21	4	59	15	110	5	32
AM	-	931	18	1	55	11	101	4	28
PP	Post procedure	801	16	1	51	10	92	3	25
SV	Subsequent visits	433	8	4	24	5	41	3	18
ECG	ECG reports	291	16	5	44	17	48	6	24
PAA	-	567	11	1	32	6	58	1	10
LIN	-	510	15	0	43	11	71	5	29
RB	-	418	8	1	17	4	46	0	6
RAD	Radiology reports	248	19	1	24	7	58	6	24
LE	Limited evaluation	140	4	2	20	5	16	1	10
SUP	Supplements	116	4	1	14	4	10	1	5
Total Cases		**1416**	**23**	**5**	**80**	**17**	**137**	**7**	**47**

One limitation of our sublanguage analysis is that we used an existing cohort with postsurgical complications identified. We have noticed the annotation of the cohort is quite noisy and some of the false postsurgical complications are actually true cases. The inequality metrics obtained here may not reflect the true inequality of the likelihood. However, we can use the words/phrases identified with high inequality of likelihood to bootstrap a better annotated data set for developing advanced informatics tools for postsurgical complication detection.

Conclusion

In this study, we have investigated the use of sublanguage analysis to facilitate NLP-enabled postsurgical complication detection. The study indicates that search-based approaches may miss some cases. The sublanguage analysis results can be used as a base to develop an information extraction system or support search-based NLP approaches by augmenting search queries. There are multiple future studies planned. One is to work with subject matter experts to improve the annotation quality of the cohort through bootstrapping. The other is to explore the use of machine learning approaches as well as sublanguage-supported search-based techniques for postsurgical complication detection.

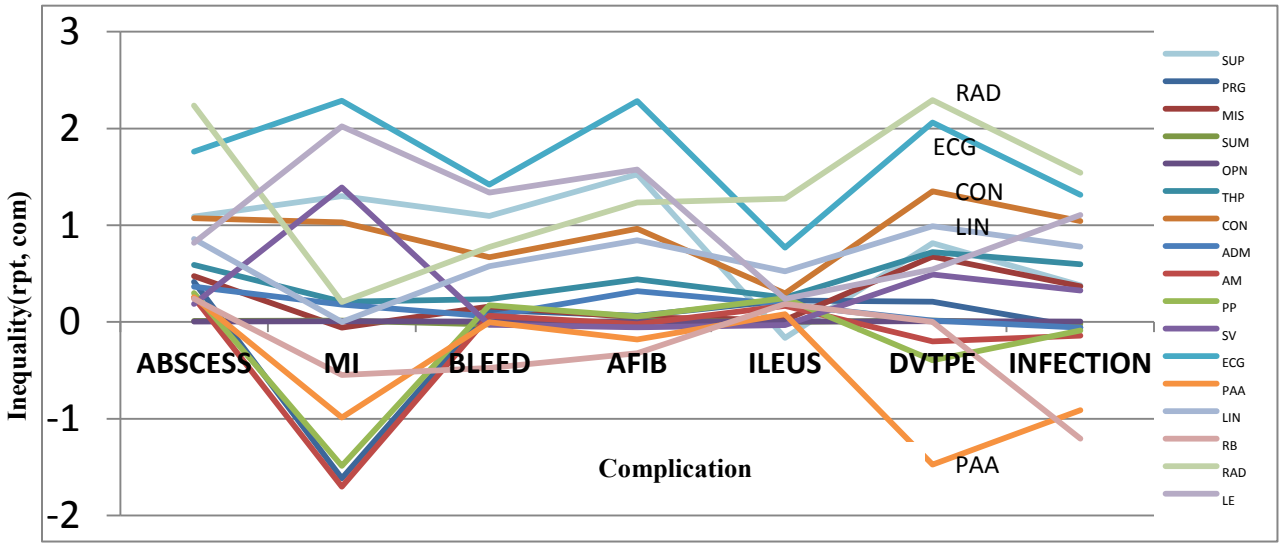


Figure 1. Inequality of report types with respect to complications

Table 4. Top 20 words/phrases identified by sublanguage analysis ranked by inequality.

Postsurgical Complication	Words/Phrases
Deep vein thrombosis (DVT)/pulmonary embolism (PE)	thrombosis; extremity edema; edema secondary; extremities bilateral; peripheral edema; wrap; negative fluid balance; left lower extremity; pe protocol; bowel ischemia; knee; popliteal vein; common femoral vein; fungal infection; sedative thromboembolism; left upper extremity; skin wound; pulmonary embolism; triglycerides
Bleeding	plasma fresh frozen; blood loss anemia; low hemoglobin; red blood cells; transfusion; transferred to icu; s hemoglobin; coagulation; gi bleed; intensive care; systolic blood pressure; hypotensive; spontaneous bacterial peritonitis; clot; hemodynamic; bilateral prophylactic mastectomy; extremity edema
Wound infections	open wound; dressing changes; wound packing; wet; vac; vacs; vacuum assisted closure; right internal jugular; bun creatinine; granulation tissue; wound care enteral nutrition; cva; keflex; wound status; granulation; fluid overload; on ventilator; wound edges; chronic pyelonephritis
Myocardial infraction	cardiac enzymes; cardiac catheterization; dominance; coronary angiography; ekg changes; cardiac monitor; st depression; ecg sinus rhythm; color flow Doppler; transthoracic echocardiogram; ischemic heart disease; coronary artery; troponin; acute myocardial infarction; coronary angiogram; metabolic acidosis; bilateral prophylactic mastectomy
Ileus	ileus; decompression; nasogastric tube; total parenteral nutrition; ng tube clamped; abdominal distention; npo; parenteral nutrition; reglan; small bowel dilatation; feels bloated; abdominal x ray; adynamic ileus; vomiting; emesis
Abscess/Leak	pressors; septic shock; resuscitated; resuscitation; sepsis; contamination; anasarca; arterial blood gas; hemodynamic instability; abdominal sepsis; mechanical ventilation; chronic pyelonephritis; hypotensive; fluid resuscitation; neo synephrine; norepinephrine; sonogram; vasopressors; vancomycin; positive pressure ventilation; pressure support

Acknowledgement

This work was made possible by joint funding from National Institute of Health grants R01LM009959A1, R01GM102282A1, and National Science Foundation grant ABI:0845523.

References

1. Leaper D, Whitaker I. Post-operative Complications: Oxford University Press; 2010.
2. Romano PS, Mull HJ, Rivard PE, et al. Validity of selected AHRQ patient safety indicators based on VA National Surgical Quality Improvement Program data. Health services research. 2009;44(1):182-204.
3. Birkmeyer JD, Shahian DM, Dimick JB, et al. Blueprint for a new American College of Surgeons: national surgical quality improvement program. Journal of the American College of Surgeons. 2008;207(5):777-82.
4. Fry DE, Pine M, Jones BL, Meimban RJ. Adverse outcomes in surgery: redefinition of postoperative complications. The American Journal of Surgery. 2009;197(4):479-84.
5. FitzHenry F, Murff HJ, Matheny ME, et al. Exploring the Frontier of Electronic Health Record Surveillance: The Case of Postoperative Complications. Medical care. 2013;51(6):509-16.
6. 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.
7. Alsara A, Warner DO, Li G, Herasevich V, Gajic O, Kor DJ. Derivation and validation of automated electronic search strategies to identify pertinent risk factors for postoperative acute lung injury. Mayo Clinic Proceedings; 2011: Elsevier; 2011. p. 382-8.
8. Sager N, Friedman C, Lyman MS. Medical language processing: computer management of narrative data. 1987.
9. Friedman C. A broad-coverage natural language processing system. Proceedings / AMIA Annual Symposium AMIA Symposium. 2000:270-4.
10. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):580-7.
11. Liu H, Bielinski SJ, Sohn S, et al. An Information Extraction Framework for Cohort Identification Using Electronic Health Records. AMIA Summits Transl Sci Proc. 2013 Mar 18;2013:149-53.
12. Liu H WS, Li D, Jonnalagadda S, Sohn S, Wagholikar K, Haug PJ, Huff SM, Chute CG Towards a semantic lexicon for clinical natural language processing Annual Symposium of American Medical Informatics Association; 2012; Chicago; 2012.

New Genetic Variants Improve Personalized Breast Cancer Diagnosis

Jie Liu, MS[1], David Page, PhD[1], Peggy Peissig, PhD[2], Catherine McCarty, PhD[3], Adedayo A. Onitilo, MD, MSCR, FACP[2,4,5], Amy Trentham-Dietz, PhD[1], and Elizabeth Burnside, MD, MPH, MS[1]

[1] University of Wisconsin, Madison, WI, US
[2] Marshfield Clinic Research Foundation, Marshfield, WI, US
[3] Essentia Institute of Rural Health, Duluth, MN, US
[4]Department of Hematology/Oncology, Marshfield Clinic Weston Center, Weston, WI, US
[5]School of Population Health, University of Queensland, Brisbane, Australia

Abstract

Recent large-scale genome-wide association studies (GWAS) have identified a number of new genetic variants associated with breast cancer. However, the degree to which these genetic variants improve breast cancer diagnosis in concert with mammography remains unknown. We conducted a case-control study and collected mammography features and 77 genetic variants which reflect the state of the art GWAS findings on breast cancer. A naïve Bayes model was developed on the mammography features and these genetic variants. We observed that the incorporation of the genetic variants significantly improved breast cancer diagnosis based on mammographic findings.

Introduction

High hopes for using genetic profiling for personalized medicine have been, in part, driven by the rapid progress of genome-wide association studies, which continue identifying more common genetic variants associated with diseases with high population prevalence. In particular, the recent Collaborative Oncological Gene-environment Study (COGS) [1], which pooled large quantities of genetic data via a massive international collaboration, more than doubled the number of known susceptibility loci that are associated to common cancers (breast, ovarian and prostate cancers). For breast cancer, over 130 institutions have collaborated and identified 41 new breast cancer associated variants [2]. One way these genetic variants could be used in clinical breast cancer care is in individualized screening recommendations and personalized diagnosis. Early attempts to incorporate genetic variants into breast cancer risk models revealed modest improvements in risk prediction accuracy. For example, adding seven SNPs to the Gail model only increased the area under the ROC curve (AUROC) from 0.607 to 0.632 [3, 4]. When ten SNPs were added to the Gail model, the AUROC increased from 0.580 to 0.618 on another dataset [5]. Incorporating these genetic variants with the mammographic findings to assess individualized risk will be highly relevant to clinical breast cancer diagnosis. In our prior study, we showed that when 22 SNPs were added to the 49 mammography features—the standard descriptors collected by radiologists on mammograms—the AUROC of the model increased from 0.693 to 0.731 [6]. This increase is statistically significant (P=0.021) [6], but the 22 SNPs only reflect the discoveries from the breast cancer GWAS up to 2010.

In this paper, we incorporated the new genetic variants and consolidated a list of 77 SNPs which reflect the state of the art of breast cancer GWAS. A great proportion of the new SNPs were contributed by COGS [2]. 41 SNPs were identified through a meta-analysis of 9 GWAS on 10,052 cases and 12,575 controls, and further showed significant association ($P<5\times 10^{-8}$) on 45,290 cases and 41,880 controls. The list also includes the 22 SNPs used in Liu et al. [6] as well as another 14 SNPs identified by several other recent studies [7-13]. We incorporated these genetic polymorphisms with the descriptors that radiologists observe on mammograms using the standardized lexicon in breast imaging, the Breast Imaging Reporting and Data System (BI-RADS). These mammography features included the shape and the margin of masses, the shape and the distribution of microcalcifications, background breast density and other associated findings. We built naïve Bayes models, using the 49 mammography features together with the 77 genetic variants. We observed that the inclusion of the genetic variants significantly improved the breast cancer diagnostic model. We discovered that the mammographic findings were more predictive for high-risk women, whereas the genetic variants were more predictive for low-risk women, which demonstrated the potential benefit of combining genetic variants and mammographic findings for personalized breast cancer diagnosis.

Data

[Subjects] The Personalized Medicine Research Project [14] at the Marshfield Clinic was used as the sampling frame to identify cases and controls. The project was reviewed and approved by the Marshfield Clinic IRB. The subjects were from a retrospective case-control design, and used in our previous study [6]. Women with a plasma

sample available, a diagnostic mammogram, and a breast biopsy within 12 months after the mammogram were included in the study. Cases were defined as women having a confirmed diagnosis of breast cancer obtained from the institutional cancer registry. Controls were confirmed through the Marshfield Clinic electronic medical records as never having had a breast cancer diagnosis by ICD-9 diagnosis code (and absence from cancer registry). Cases included both invasive breast cancer and ductal carcinoma in situ. We employed an age matching strategy to construct case and control groups that were similar in age distribution. We selected a control whose age was within five years of the age of each case. We decided to focus on high-frequency/low-penetrance SNPs that affect breast cancer risk as opposed to low frequency SNPs with high penetrance or intermediate penetrance. We excluded individuals who had a known high-penetrance genetic mutation.

[Genetic Variants] Our study included the 77 genetic variants (in Table 1) which were identified by the recent large-scale genome-wide association studies. 22 of these SNPs were evaluated in the previous study of Liu et al. (2013) [6]. Among the 55 new SNPs, 41 were identified by COGS [2], and 14 SNPs were included based on several other recent studies [7-13]. It is estimated that the current list of SNPs explains 14% of familial breast cancer risk [2].

[Mammography Features] Mammography is the most common breast cancer screening test, and the only one supported by multiple randomized trials demonstrating reduction in mortality rate [15]. There is a long history of development and codification of features observed by radiologists on mammograms. The American College of Radiology developed the BI-RADS lexicon to standardize mammographic findings and recommendations. The BI-RADS lexicon consists of 49 descriptors, including the characteristics of masses and microcalcifications, background breast density and other associated findings. Mammography data was historically recorded as free text reports in the electronic health record, and thus it was difficult to directly access the information contained therein. We used a parser to extract these mammography features from the text reports; the parser was shown to outperform manual extraction [16, 17]. After extraction, each mammography feature took the value "present" or "not present" except that the variable *mass size* was discretized into three values, "not present", "small" and "large", depending on whether there was a reported mass size and whether any dimension was larger than 30mm.

A BI-RADS assessment category was assigned to each mammogram by the interpreting radiologist, which indicated the radiologist's assessment of the absence or presence of breast carcinoma. In our study, the BI-RADS assessment category took values, with an order of increasing probability of malignancy, of 1, 2, 3, 0, 4a, 4, 4b, 4c and 5. We used the BI-RADS assessment category as the predictions from the radiologists. Our study only included diagnostic mammograms, and all the screening mammograms were excluded. For cases, we selected the mammograms within one year prior to diagnosis. For controls, we selected the mammograms within one year prior to biopsy. If there were multiple diagnostic mammograms during that one year time period, we selected the mammogram with a more suspicious BI-RADS category, with subsequent tiebreakers being recency and the number of extracted features.

Model

We built breast cancer diagnosis models using Naïve Bayes, which can be regarded as the weighted average of risk factors. Naive Bayes assumes that all features are conditionally independent of one another given the class [18]. Although this assumption seems strong, it generally works well in practical problems and provides easy interpretation of the risk contribution from different factors. In our experiments, we used the Naïve Bayes implementation in WEKA [19].

In total, we constructed three types of models on different sets of features. The first model was built purely on the 49 mammography features, namely the *Breast Imaging model*. The second type of model was based purely on genetic variants, namely the genetic models. Since we would like to align our study with previous work, we tested three sets of genetic variants. The first set consisted of the 10 SNPs in the study of Wacholder et al. (2010) [5]. The second included the 22 SNPs in the study of Liu et al. (2013) [6]. The last set was our full list of the 77 SNPs. We denote the three genetic models as *Genetic-10*, *Genetic-22* and *Genetic-77* models. The third type of model was built on the 49 mammography features and the genetic variants together, namely the *combined models*. Since we had three sets of genetic variants with different sizes, we had three combined models, namely *Combined-10*, *Combined-22* and *Combined-77* models. In both the genetic models and the combined models, we handled the genetic variants in the following way rather than using original genotypes of each SNP. We only introduced one additional variable, the total count of risky alleles the person carries in the DNA. This way of coding genetic variants was used in several models such as [5], and is helpful to build risk models when each SNP only has a small contribution to the risk.

We treated the BI-RADS category scores from the radiologists as the predictions from the radiologists, namely the *baseline clinical assessment*. We constructed ROC curves for each model, and used the area under the curve (AUC) as a measure of performance. We also provided the precision-recall (PR) curves for the models. We evaluated the models using 10-fold cross-validation.

Table 1. The 77 SNPs identified to be associated to breast cancer.

SNP	Chr	Ref	Notes[1]	SNP	Chr	Ref	Notes
rs11249433	1	[20]	WL	rs2380205	10	[12]	
rs616488	1	[2]		rs10995190	10	[12]	
rs1045485	2	[21]	GWL	rs704010	10	[12]	
rs17468277	2	[22]	L	rs2981579	10	[12]	
rs4666451	2	[23]	L	rs7072776	10	[2]	
rs13387042	2	[20, 24]	GWL	rs7904519	10	[2]	
rs4849887	2	[2]		rs11199914	10	[2]	
rs2016394	2	[2]		rs11814448	10	[2]	
rs1550623	2	[2]		rs2107425	11	[23]	L
rs16857609	2	[2]		rs3817198	11	[20, 23]	GWL
rs4973768	3	[25]	L	rs614367	11	[12]	
rs6762644	3	[2]		rs3903072	11	[2]	
rs12493607	3	[2]		rs11820646	11	[2]	
rs9790517	4	[2]		rs6220	12	[26, 27]	L
rs6828523	4	[2]		rs1292011	12	[10, 13]	
rs10941679	5	[20, 28]	WL	rs17356907	12	[2]	
rs30099	5	[23]	L	rs10771399	12	[2]	
rs889312	5	[23]	GWL	rs12422552	12	[2]	
rs981782	5	[23]	L	rs11571833	13	[2]	
rs1353747	5	[2]		rs999737	14	[20]	WL
rs1432679	5	[2]		rs2236007	14	[2]	
rs10069690	5	[2]		rs2588809	14	[2]	
rs10472076	5	[2]		rs941764	14	[2]	
rs2046210	6	[29]	L	rs3803662	16	[20, 23, 24]	GWL
rs2180341	6	[30]	L	rs8051542	16	[23]	L
rs17530068	6	[9]		rs12443621	16	[23]	L
rs3757318	6	[12]		rs13329835	16	[2]	
rs11242675	6	[2]		rs17817449	16	[2]	
rs204247	6	[2]		rs6504950	17	[25]	L
rs720475	7	[2]		rs1436904	18	[2]	
rs13281615	8	[20, 23]	GWL	rs527616	18	[2]	
rs9693444	8	[2]		rs8170	19	[8]	
rs11780156	8	[2]		rs4808801	19	[2]	
rs6472903	8	[2]		rs3760982	19	[2]	
rs2943559	8	[2]		rs2284378	20	[9]	
rs1011970	9	[12]		rs2823093	21	[10]	
rs865686	9	[7, 11, 13]		rs132390	22	[2]	
rs10759243	9	[2]		rs6001930	22	[2]	
rs2981582	10	[20, 23, 28, 30, 31]	GWL				

Results

We identified 362 cases and 377 controls. Among the cases, there were 358 Caucasians, three non-Caucasians and one case whose race information was unknown. Among the controls, there were 373 Caucasians and four non-Caucasians. We do not disclose the race/ethnicity information of these non-Caucasians for privacy concerns. Subject characteristics including age distribution and family history of breast cancer are described in Table 2. There were more young people (age <50) in the case group than in the control group, and the proportion of elderly people (age ≥65) was roughly the same in the case group and in the control group. For the family history of breast cancer, we observed a considerable larger proportion of people with family history in the case group (45.3%) than in the control group (33.7%), which demonstrated the family aggregation of breast cancer.

[1] G stands for being used in the study by Gail (2008, 2009) [3, 4]; W stands for being used in the study by Wacholder et al. (2010) [5]; L stands for being used in the study of Liu et al. (2013) [6].

Table 2. The distribution of age and family history of breast cancer.

	Cases	Controls	All		Cases	Controls	All
Age Group				**Family History**			
<50	81 (22.4%)	58 (15.4%)	139 (18.8%)	**Yes**	164 (45.3%)	127 (33.7%)	291 (39.4%)
50-65	123 (34.0%)	168 (44.6%)	291 (39.4%)	**No**	188 (51.9%)	236 (62.6%)	424 (57.4%)
≥65	158 (43.6%)	151(40.0%)	309 (41.8%)	**N/A**	10 (2.8%)	14 (3.7%)	24 (3.2%)

The Performance of the Three Combined Models

The ROC and the PR curves for the baseline clinical assessment, the Breast Imaging model and the three combined models are provided in Figure 1. For each model, we vertically average [32] the ROC curves from the ten replications of the 10-fold cross-validation to obtain the final curve; we do likewise for the PR curves. The area under the ROC curves for the Breast Imaging model, the Combined-10 model, the Combined-22 model and the Combined-77 model are 0.693, 0.712, 0.733 and 0.760. The ROC curve of the Combined-77 model almost completely dominates the ROC curve of the Breast Imaging model, which suggests that the 77 genetic variants can help to improve breast cancer diagnosis based on mammographic findings. We perform a two-sided paired *t*-test on the area under the ten ROC curves of the Breast Imaging model and the area under the ten ROC curves of the combined model from the 10-fold cross-validation, and the difference between them is significant with a *P*-value 0.00047. We further compare the AUROC of the Combined-77 model and the Combined-22 model with a two-sided paired *t*-test, and the difference between them is significant with a *P*-value 0.0046, which demonstrates the discriminative power of the 55 recently identified SNPs. From PR curves, we note that the combined models dominate the Breast Imaging model and the baseline clinical assessment in the high recall region (>0.8) in which clinicians operate, and therefore we want to optimize.

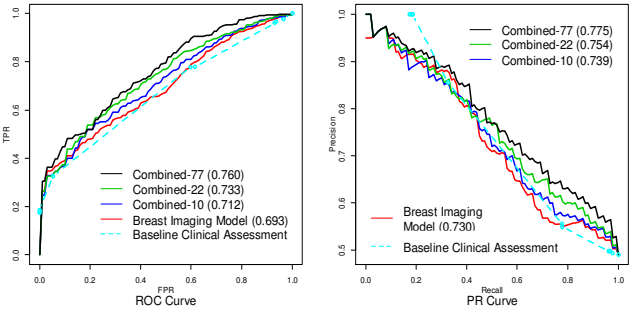


Figure 1. The ROC curves and PR curves for the baseline clinical assessment, the Breast Imaging model the three combined models.

The Performance of the Three Genetic Models

Furthermore, we compare the discriminative power of the three genetic models, namely the Genetic-10 model, the Genetic-22 model and the Genetic-77 model. The ROC curves and the PR curves for the three genetic models are provided in Figure 2, respectively. For each model, we vertically average the curves from the 10-fold cross-valuation to obtain the final curve. The area under the ROC curves for the Genetic-10 model, the Genetic-22 model and the Genetic-77 model are 0.591, 0.622 and 0.684, which demonstrates that the more associated SNPs the genetic model includes, the more discriminative the model becomes. We also use a two-sided paired *t*-test to compare the area under the ROC curves yielded by the three genetic models. The Genetic-77 model outperforms both the Genetic-22 model (*P*=0.028) and the Genetic-10 model (*P*=0.0068).

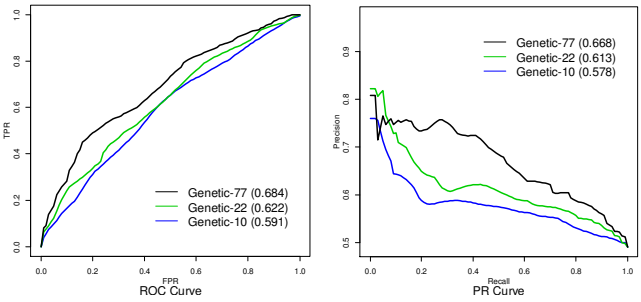


Figure 2. The ROC and PR curves for the three genetic models.

Comparing Breast Imaging and Genetic-77 Model

We compare the performance of the Breast Imaging model, the Genetic-77 model and the Combined-77 model. The corresponding ROC curves and the PR curves for the three models are shown in Figure 3. We observe that the mammography features are more predictive for women with a high probability of cancer (low FPR region in ROC space) whereas genetic variants are more predictive for women with a low probability of cancer (mid/high FPR region in ROC space). Note that the Genetic-77 model describes

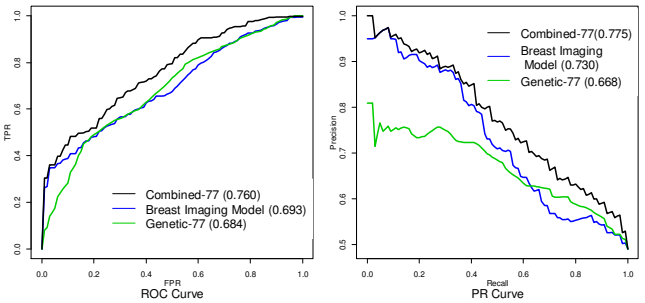


Figure 3. The ROC curves and PR curves for the Breast Imaging model, the Genetic-77 model and the Combined-77 model.

the patient's inherited breast cancer risk in DNA. However, after the patient starts developing malignant features on mammograms, mammographic findings (Breast Imaging model) provide superior discrimination. Still, knowing the genetic information can further improve the accuracy of breast cancer diagnosis even at higher baseline risk.

Discussion

The primary contribution of our study is to show that the genetic variants can significantly improve breast cancer diagnosis on mammographic findings, resulting in reduced false positives and alleviated risk of overdiagnosis. This result indicates promise for translating discoveries from massive collaborative GWAS into clinical breast cancer diagnosis. Our study includes the most up-to-date breast cancer associated SNPs, the majority identified and/or verified through the massive COGS (over 55k cases and over 54k controls), and therefore these new SNPs are credible and can explain a larger proportion of familial breast cancer risk. Indeed, we observe that the Combined-77 model significantly outperforms the Combined-22 model used in our previous study [6]. We also demonstrate that the Genetic-77 model significantly outperforms the Genetic-22 model. The increased discriminative power derived from the new 55 SNPs identified by recent published studies [2] highlights the rapid progress the breast cancer GWAS community has made since 2010. Furthermore, we make a novel discovery that mammography features are more predictive for high-risk women whereas genetic variants are more predictive for low-risk women, which explains the benefit of combining genetic variants and mammographic findings for personalized breast cancer diagnosis.

Our study, in a novel way, differs from the previous study of Wacholder et al. (2010) [5] which adds ten genetic variants to the Gail model, a risk model based on self-reported demographic and personal risk factors. The unique contribution in our study is that we include mammography features which represent richer phenotypic data directly relevant to breast cancer diagnosis and thus provide high signal. Therefore, our study contributes the potential clinical impact of translating exciting discoveries from GWAS to the patient experience at diagnosis. The additional discriminative power from these genetic variants can significantly rule out the false positives of mammogram screening, and therefore has the potential to decrease recommendations for unnecessary breast biopsies. Of course, it will be interesting to combine the epidemiology features in Gail model, the mammography features and the SNPs for more accurate personalized breast cancer diagnosis.

Limitations of our study include small sample size and the pitfalls of data extraction from text reports. We understand that parsing mammography features from text reports may introduce noise into the data. However, despite the challenges inherent in extracting accurate data, which may affect our results, we are encouraged that improvements in predictive accuracy remain, especially after observing the discriminative power of genetic factors alone in the genetic models. Furthermore, we recognize that methodological issues in our study may represent shortcomings but also signify opportunities for future investigation. First, we do not explicitly model how individual SNPs function to alter breast cancer risk, nor do we model potential SNP interactions [33]. Our current model only adds one extra feature which simply counts the totally number of risky alleles, assuming that the effect size of the genetic variants are the same and that the genetic effect of the genetic variants is non-mechanistic and additive. We do not model the individual SNPs for the curse of dimensionality concern; each individual SNP only confers a fairly mild relative risk and if we model them individually, the model will perform poorly on test data unless a larger cohort of training data is available. Modeling SNP-SNP interaction is even harder and requires more training data.

Second, we do not differentiate the different subtypes of breast cancers (for example, the estrogen-receptor status and progesterone-receptor status) in the current study. Breast cancer is a complex and heterogeneous disease with different subtypes, including two main subtypes of estrogen receptor (ER) negative tumors (basal-like and human epidermal growth factor receptor-2 positive/ER− subtype) and at least two types of ER positive tumors (luminal A and luminal B) [34, 35]. These molecular subtypes are important predictors of breast cancer mortality [36] and have different genetic susceptibility [37]. Therefore it is desirable to tease them apart in the pursuit of increasingly personalized breast cancer care.

Nevertheless, we are encouraged by these promising results in our current study, especially after the disappointment [38] and caution [39] in the early years of translating GWAS discoveries to personalized risk prediction. We hope that the rapid progress being made through these massive collaborative studies together with our growing knowledge about breast cancer mechanisms and genotype-phenotype relationships will bring us even closer to the practical personalized breast cancer diagnosis and treatment.

Acknowledgements

The authors acknowledge the support of the Wisconsin Genomics Initiative, NCI grant R01CA127379-01 and its ARRA supplement 3R01CA127379-03S1, NIGMS grant R01GM097618-01, NLM grant R01LM011028-01, NIEHS grant 5R01ES017400-03, the UW Institute for Clinical and Translational Research (ICTR) and the UW Carbone Cancer Center.

Reference

1. Bahcall, O.G., *iCOGS collection provides a collaborative model. Foreword.* Nat Genet, 2013. **45**(4): p. 343.
2. Michailidou, K., et al., *Large-scale genotyping identifies 41 new loci associated with breast cancer risk.* Nat Genet, 2013. **45**(4): p. 353-61, 361e1-2.
3. Gail, M.H., *Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model.* J Natl Cancer Inst, 2009. **101**(13): p. 959-63.
4. Gail, M.H., *Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk.* J Natl Cancer Inst, 2008. **100**(14): p. 1037-41.
5. Wacholder, S., et al., *Performance of common genetic variants in breast-cancer risk models.* N Engl J Med, 2010. **362**(11): p. 986-93.
6. Liu, J., et al. *Genetic Variants Improve Breast Cancer Risk Prediction on Mammograms*. in *American Medical Informatics Association Symposium*. 2013.
7. Warren, H., et al., *9q31.2-rs865686 as a susceptibility locus for estrogen receptor-positive breast cancer: evidence from the Breast Cancer Association Consortium.* Cancer Epidemiol Biomarkers Prev, 2012. **21**(10): p. 1783-91.
8. Stevens, K.N., et al., *19p13.1 is a triple-negative-specific breast cancer susceptibility locus.* Cancer Res, 2012. **72**(7): p. 1795-803.
9. Siddiq, A., et al., *A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11.* Hum Mol Genet, 2012. **21**(24): p. 5373-84.
10. Ghoussaini, M., et al., *Genome-wide association analysis identifies three new breast cancer susceptibility loci.* Nat Genet, 2012. **44**(3): p. 312-8.
11. Fletcher, O., et al., *Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study.* J Natl Cancer Inst, 2011. **103**(5): p. 425-35.
12. Turnbull, C., et al., *Genome-wide association study identifies five new breast cancer susceptibility loci.* Nat Genet, 2010. **42**(6): p. 504-7.
13. Antoniou, A.C., et al., *A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population.* Nat Genet, 2010. **42**(10): p. 885-92.
14. McCarty, C., et al., *Marshfield Clinic Personalized Medicine Research Project (PMRP): design, methods and recruitment for a large population-based biobank.* Personalized Med, 2005. **2**: p. 49-79.
15. Marmot, M., et al., *The benefits and harms of breast cancer screening: an independent review.* British Journal of Cancer, 2013. **108**(11): p. 2205--2240.
16. Houssam, N., et al., *Information Extraction for Clinical Data Mining: A Mammography Case Study*, in *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*. 2009, IEEE Computer Society.
17. Percha, B., et al., *Automatic classification of mammography reports by BI-RADS breast tissue composition class.* J Am Med Inform Assoc, 2012. **19**(5): p. 913-6.
18. Lowd, D. and P. Domingos. *Naive Bayes models for probability estimation*. in *Proceedings of the 22nd international conference on Machine learning*. 2005.
19. Hall, M., et al., *The WEKA data mining software: an update.* SIGKDD Explor. Newsl., 2009. **11**(1): p. 10--18.
20. Thomas, G., et al., *A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1).* Nat Genet, 2009. **41**(5): p. 579-84.
21. Cox, A., et al., *A common coding variant in CASP8 is associated with breast cancer risk.* Nat Genet, 2007. **39**(17293864): p. 352-358.
22. Odefrey, F., et al., *Common genetic variants associated with breast cancer and mammographic density measures that predict disease.* Cancer Res, 2010. **70**(20145138): p. 1449-1458.

23. Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(17529967): p. 1087-1093.
24. Stacey, S.N., et al., *Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.* Nat Genet, 2007. **39**(7): p. 865-9.
25. Ahmed, S., et al., *Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2.* Nat Genet, 2009. **41**(19330027): p. 585-590.
26. Biong, M., et al., *Genotypes and haplotypes in the insulin-like growth factors, their receptors and binding proteins in relation to plasma metabolic levels and mammographic density.* BMC Med Genomics, 2010. **3**(20302654): p. 9-9.
27. Kelemen, L.E., T.A. Sellers, and C.M. Vachon, *Can genes for mammographic density inform cancer aetiology?* Nat Rev Cancer, 2008. **8**(18772892): p. 812-823.
28. Stacey, S.N., et al., *Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer.* Nat Genet, 2008. **40**(18438407): p. 703-706.
29. Zheng, W., et al., *Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1.* Nat Genet, 2009. **41**(19219042): p. 324-328.
30. Gold, B., et al., *Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33.* Proc Natl Acad Sci U S A, 2008. **105**(18326623): p. 4340-4345.
31. Hunter, D.J., et al., *A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.* Nat Genet, 2007. **39**(17529973): p. 870-874.
32. T, F., *An introduction to ROC analysis.* Pattern Recognition Letters, 2006. **27**(8): p. 861--874.
33. Turnbull, C., et al., *Gene-gene interactions in breast cancer susceptibility.* Hum Mol Genet, 2012. **21**(4): p. 958-62.
34. Carey, L.A., et al., *Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study.* Jama, 2006. **295**(21): p. 2492-502.
35. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
36. Haque, R., et al., *Impact of breast cancer subtypes and treatment on survival: an analysis spanning two decades.* Cancer Epidemiol Biomarkers Prev, 2012. **21**(10): p. 1848-55.
37. Garcia-Closas, M., et al., *Genome-wide association studies identify four ER negative-specific breast cancer risk loci.* Nat Genet, 2013. **45**(4): p. 392-8, 398e1-2.
38. Goldstein, D.B., *Common genetic variation and human traits.* N Engl J Med, 2009. **360**(17): p. 1696-8.
39. Kraft, P. and D.J. Hunter, *Genetic risk prediction--are we there yet?* N Engl J Med, 2009. **360**(17): p. 1701-3.

Phonetic Spelling Filter for Keyword Selection in Drug Mention Mining from Social Media

Pranoti Pimpalkhute, MS[1], Apurv Patki, MS[1], Azadeh Nikfarjam, Phd[2], Graciela Gonzalez, Phd[2]

[1]Arizona State University, Department of Computer Science, Tempe, AZ;
[2]Arizona State University, Department of BioMedical Informatics, Scottsdale, AZ

Abstract

Social media postings are rich in information that often remain hidden and inaccessible for automatic extraction due to inherent limitations of the site's APIs, which mostly limit access via specific keyword-based searches (and limit both the number of keywords and the number of postings that are returned). When mining social media for drug mentions, one of the first problems to solve is how to derive a list of variants of the drug name (common misspellings) that can capture a sufficient number of postings. We present here an approach that filters the potential variants based on the intuition that, faced with the task of writing an unfamiliar, complex word (the drug name), users will tend to revert to phonetic spelling, and we thus give preference to variants that reflect the phonemes of the correct spelling. The algorithm allowed us to capture 50.4 – 56.0 % of the user comments using only about 18% of the variants.

Keywords: Information Retrieval, Natural Language Processing and Free Text Data Mining, Spelling-Error

Introduction

The question to ask in information extraction from social media postings is not whether valuable information is present in the user data, but how to find it among the millions of daily postings and how to work around the limitations that these sites necessarily impose on automatic requests.

Social networking postings can indeed be a treasure trove of data. Twitter alone observes around 58 million tweets(3) per day. Obtaining the right ones might be tricky, however, even when using the site-provided APIs (Application Programming Interface) for data collection. For example, Twitter provides Streaming and Search APIs to collect tweets, but in order to collect tweets for a particular topic, appropriate keywords should be first selected and given to the API. Twitter allows to track up to 400 keywords per application key, returning all matching Tweets up to a volume equal to the streaming cap (which is about 1% of the totality of all public streamed tweets). The GooglePlus API also restricts calls up to 1000 requests per day.

Our ongoing work(1)(2) to extract mentions of adverse reactions of drugs directly from patient comments posted on social networks has exposed the significance and nuances of a common problem: medical terms, and specifically, drug names are particularly difficult for users to spell correctly, and frankly, they usually make no obvious effort to do so when posting a message online. Thus, given that for automatic collection of postings related to the drugs the drug name is the keyword used to obtain the postings, including misspelled versions of the drug name as keywords is important. Consider the following examples of Tweets obtained for various spellings of *Seroquel*:

- @PsychoIogicaI HA! Not if you're on # **Seroquil** . EXTREMELY vivid dreams that stay in conscious memory. Very # Freaky ! Any idea why?
- @BipolarBlogger did you ever try the **Seriquel** XR??? It has a less sedative effect and has a longer lasting effect
- Gone from 50mg to 150mg of **Serequel** last night. Could barely wake up this morning and I feel like my body is made of lead
- @AndrewH_Smith Is the Inderal helpful? And yeah, they are short lasting but non addictive. You could try **Seraquel** too but it's pretty strong

However, algorithms to generate word variants (using 1 or 2 edit distance and typographical –keyboarding- errors, for example) produce in excess of 300 or more variants per drug name of average length. If we consider that the total number of drugs currently listed in DrugBank(4) website is around 6800, about 2 million keywords to track postings related to all drugs would be necessary, exceeding the limit imposed for an instance of the Twitter Streaming API crawler with only 2 or 3 drugs. Even if this limit could be bypassed via multiple application instances, handling and deploying such a large number of query terms might be impractical and unnecessary, as many misspellings are not common enough to warrant monitoring. On the other hand, without including the common misspellings, many postings would be missed. In fact, the number of postings that use the most common misspellings of drug names often exceed those that use correct spelling. Thus, we are faced with the problem of generating misspelled variants for a drug name and then filtering them to select the most common ones in order to remain within the crawling API limitations when mining drug-related postings in social media.

This paper describes a method to generate most probable misspelled drug name variants for querying social media postings. The method is based on the intuitive notion that people will tend to spell drug names phonetically, the "default" used by young children when spelling an unfamiliar word. Including these most probable misspelled variants allows us to collect valuable posts from social networking sites that we would have missed otherwise. There has been some prior work to this effect, but in general, most focus on correcting or detecting misspellings, not generating them. For example, Senger, et al.,(6) proposed an auto-correction algorithm to prevent errors in drug spelling. The web site Drugs.com allows a user to type a drug name phonetically (These approaches assume that all the text is available and then apply algorithms for spelling correction, on the contrary, in our task we have to generate keywords first to crawl the data from social media. An example of the later includes an approach to generate spelling variants for proper nouns, proposed by Bhagat and Hovy(5), in order to detect names of foreign places and people published after transliteration. Spelling mistakes of drug names can occur because of pronunciation error and typing errors. Hence it is important to consider both of these error types while generating probable misspelled versions of a drug name... Thus, our task is to generate a balanced list of keywords that can give maximum coverage (extracting a good portion of the useful comments) from social media sites. The rest of this paper covers the methods, evaluation, and results used for this task. For ease of reference, we will refer to the Twitter Streaming API, GooglePlus API and Facebook API collectively as the "crawling API". Small variations in the APIs themselves are not relevant to the task.

Methods

Considering the possible ways the misspellings may occur in drug names, we sought to develop an algorithm to generate a list of all the likely misspelled variants of a drug name based on a simple 1-edit distance algorithm, and then filtering it using phonetic spelling. We evaluated the approach as to its ability to generate a list with maximum coverage of social networking postings for a minimum list size. For evaluation purposes, we thoroughly examined social network postings for 4 drugs – Paxil, Prozac, Seroquel and Olanzapine. The number of tweets collected directly from the user interface for Twitter for Paxil were 334 using 18 variants, for Prozac were 186 using 18 variants, for Seroquel 146 using 17 variants and for Olanzapine were 89 using 15 variants.

Tools and Dataset. We utilize three different social media resources in our system: Facebook, Twitter and GooglePlus. There are many phonetic spelling algorithms available. We choose to use the LOGIOS Lexicon Tool (7) and Metaphone library (8,9) as they are one of the most common APIs. We used these libraries to get variants with similar pronunciation for a drug name. The LOGIOS Lexicon Tool generates a list of different pronunciations by expanding the original word into machine readable pronunciation which is encoded using the modified form of Arpabet system(10).

The Metaphone phonetic algorithm is an improvement over the Soundex phonetic algorithm, where the words are encoded to the same representation so that they can be grouped despite minor differences. For example, Table 1 shows the encodings of words using the CMU pronunciation (11) and the Metaphone library. The words with similar pronunciations of "Prozac" obtained by the CMU library are "Prozak" and "Proxac", whereas the similar pronunciation words obtained from Metaphone encoding are "Przac" and "Prozak".

T**able 1.** CMU and Metaphone encoding

Word Variants	**CMU Expanded Pronunciation**	**Metaphone Encoding**
PROZAC*	***P R OW Z AE K***	***PROZACPRSK***
POZAC	P AA Z AH K	POZACPSK
PRZAC	P R Z AE K	**PRZACPRSK**
PROAC	P R OW AE K	PROACPRK
PROZAK	**P R OW Z AE K**	**PROZAKPRSK**
PROXAC	**P R OW Z AE K**	PROXACPRKS

*correct spelling of the word

Figure 1 shows the flowchart for the method used in this paper. The first step is to generate all variants of a word within 1-edit distance (Levenshtein distance, words that vary from the original by a single-character insertion, deletion, or substitution). The number of variants at even 1-edit distance to a drug name are very large and the count shoots up for 2 or more edit distances. Consider the drug "Paxil": with only 5 characters in length, the number of variants obtained with 1-edit distance are 238. For the drug "Olanzapine" with length of 10 characters, there are 503 words within 1-edit distance. Table 2 shows the number of variants of drug names obtained by just 1-edit distance.

Moving further, in order to compare the misspelled variants to the original word in pronunciation, we applied the CMU pronunciation and Metaphone algorithm to find those having the same pronunciation originating from the different spellings. For example, "Prozac" and "Prozak" have the same pronunciation. The results obtained from both libraries were useful, as they resulted in a significant number of social network postings, so we couldn't right out eliminate any of them. For example, for the words "Paxil" and "Paxcil", CMU lists the same pronunciation, whereas Metaphone does not, while Metaphone considers "Paxil" and "Paxial" as having similar pronunciation, and CMU does not. However, just combining the variants obtained from the two algorithms still results in a very large number of words, considering the limitations of the crawling API. For example, for the word "Paxil", the number of variants in the combined list is 85 words (those with the same pronunciation as the original word). Table 2 shows the number of words obtained from the CMU and the Metaphone libraries, as well as the combination of both.

Thus, phonetic pronunciation alone, although it reduces the list by a third, might not be sufficient if one wishes to monitor more than a handful of drugs. To find out which of the given variants are common misspellings, we used the Google custom search API(12), issuing a query per variant composed of each of the selected misspelled variants that have the same pronunciation as the original drug name plus the word 'drug' (to reduce the "noise" generated by pages referring to other topics). Google hits were used as an estimate of how prevalent the misspelled word would be on the social networks. We then ranked the words according to the number of hits, and choose the top k, setting k to the rank where the rate of change of Google hits drops significantly. Using this threshold, the list of highly probable misspelled variants was set to the top k (18 for Paxil, as shown in Table 2).

The lists generated by the algorithm were used to obtain comments by users of the three social networking sites mentioned before (Twitter, Facebook, and Google Plus).

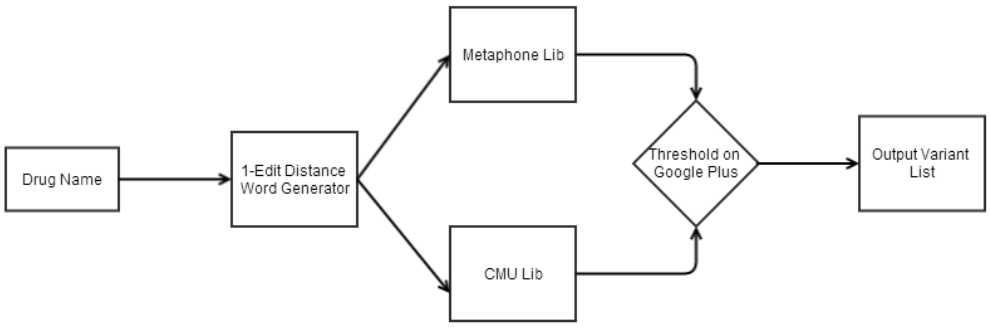


Figure 1. Control Flow Chart

Table 2. Statistics of misspelled variants.

	Paxil	Prozac	Seroquel	Olanzapine
Levenshtein (1-edit) distance words	238	291	397	503
CMU lib words generated	21	18	27	31
Metaphone words generated	79	103	121	338
Combining the two lists	85	104	119	327
Keywords selected by proposed algorithm	18	18	17	15

Evaluation Method 1. We used four drugs for the purpose of experiments. In order to evaluate our proposed approach, we want to compute the fraction of useful posts that the method was able to capture from the crawling API using the variants generated by the algorithm. Since it is not possible to retrieve the exact number of posts corresponding to a variant from the API due to its limitations, we retrieved the comments using screen scraping in order to get a true measure of misspelled variants. For the evaluation of this algorithm we used coverage of the sampled list as an evaluation metric. The coverage of sampled list can be defined as,

$$Comments\ Coverage\ (\alpha) = \frac{Number\ of\ tweets\ from\ sampled\ list}{Number\ of\ \ tweets\ from\ complete\ list}$$

$$Keywords\ Coverage\ (\beta) = \frac{Number\ of\ keywords\ selected}{Total\ number\ of\ keywords\ generated}$$

where the sampled list is the list selected from our algorithm and complete list is combined output of the CMU and Metaphone libraries. Comments Coverage give us fraction of tweets the method was able to capture using the sampled list. The metric "keywords coverage" evaluates the fraction of keywords used as tracking words.

Evaluation Method 2. In this evaluation strategy we compared the results of our algorithm to a random keyword selector which is our baseline. We show that our algorithm produces a significant improvement over random keyword selector. The essence of our method is to capture a large amount of relevant data from a small variant list. Figure 3 show number of Google Plus comments and number of tweets collected respectively for random sample of variants and the variants sorted by Google Custom Search API.

Results

Evaluation Method 1. Figure 2 show the plots for number of comments obtained from Twitter vs the number of drug variants. The X-axis represents the combined list of drug variants obtained from the CMU pronunciation and the Metaphone library and the Y-axis represents the Google hits for the variant. The drug variants were ranked according to the Google hits obtained from the custom search API. It is evident from Figure 2 that total number of comments does not increase by much after a particular point. The algorithm allowed us to capture 50.42 – 55.97% of comments using about 18.29% of the variants.

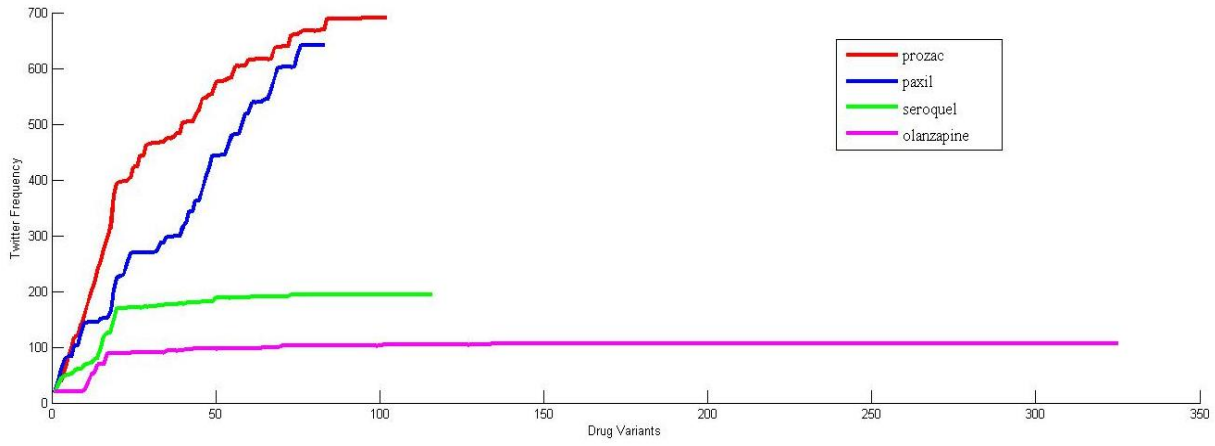


Figure 2. Number of Tweets vs Drug Variants.

Evaluation Method 2. Figure 3 shows that our method has an advantage when the keyword coverage is less, and that the method can capture useful data with minimal keyword coverage. For instance, for 20% keyword coverage the random selector captured 32 tweets while our approach captured 170 tweets for Seroquel.

Table 4. Evaluation for Twitter and GooglePlus

Drug Name	Twitter Comments Coverage	GooglePlus Comments Coverage	Keyword Coverage
Prozac	45.44138929	52.7607362	17.30769231
Paxil	25.85669782	54.54545455	21.17647059
Seroquel	65.28497	62.29508	14.40678
Olanzapine	65.09433962	54.28571429	4.573170732
Average	50.417	55.971	

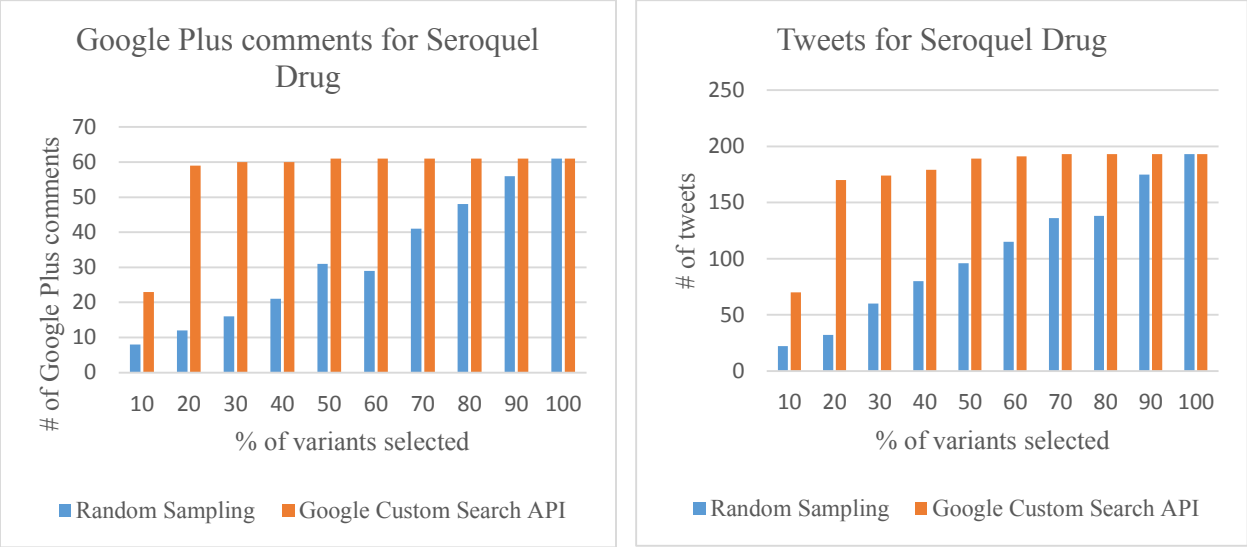


Figure 3. Google Plus comments and Tweets for Custom Search API sorted variants and Random variants.

Discussion

Our work focuses on finding a balance between the restrictions imposed by the crawling APIs and the many variants of a drug name needed to capture the large amount of data that hides behind these restrictions. Other approaches seek to correct misspelling by mapping the drug names from free text to standard nomenclature(13), but given the context of this work, these methods will fail in extracting data from social media given that all the data cannot be captured beforehand and then filtered out.

The main limitation of this algorithm is that some common misspellings that are due to typographical errors could be missed and might be commonly used. The false negative rate cannot be adequately computed since the universal set is not known. Moreover, the data obtained from social networking sites is complicated. For example, for the word *Prozac,* there are tweets that are not related to the drug "*Local woodstock continues After the wild cats and a live connection with Ibiza now are the* ***prozec*** *mckenzie on stage ... Tonight only".* The keyword coverage can be manipulated to achieve a higher comment coverage, adjustin for the number of drugs to track and API limits. Moreover, it is important to appreciate that crawling API is a resource which can be used in a better way if we have tracking words that are relevant.

References

1. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annu. Symp. Proc. 2011 Jan;2011:1019–26.
2. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G. Towards Internet-Age Pharmacovigilance : Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. 2010;(July):117–25.
3. Twitter Statistics [Internet]. Available from: http://www.statisticbrain.com/twitter-statistics/
4. DrugBank Statistics [Internet]. Available from: http://www.drugbank.ca/stats
5. Bhagat R, Hovy E, Way A, Rey M Del. Phonetic Models for Generating Spelling Variants.
6. Senger C, Kaltschmidt J, Schmitt SPW, Pruszydlo MG, Haefeli WE. Misspellings in drug information system queries: characteristics of drug name spelling errors and strategies for their prevention. Int. J. Med. Inform. Elsevier Ireland Ltd; 2010 Dec [cited 2013 Oct 5];79(12):832–9.
7. LOGIOS Lexicon Tool [Internet]. Available from: http://www.speech.cs.cmu.edu/tools/lextool.html
8. Metaphone Wiki [Internet]. Available from: http://en.wikipedia.org/wiki/Metaphone
9. Soundex Wiki [Internet]. Available from: http://en.wikipedia.org/wiki/Soundex
10. Arpabet System [Internet]. Available from: http://en.wikipedia.org/wiki/Arpabet
11. CMU Pronouncing Dictionary [Internet]. Available from: http://en.wikipedia.org/wiki/CMU_Pronouncing_Dictionary
12. Google Custom Search API. Available from: https://developers.google.com/custom-search/
13. Levin MA, Krol M, Ph D, Doshi AM, Reich DL. Extraction and Mapping of Drug Names from Free Text to a Standardized Nomenclature AMIA 2007 Symposium Proceedings Page - 438 AMIA 2007 Symposium Proceedings Page - 439. 2007;438–42.

Appendix

Table Top k drug name variants and their corresponding Google Hits obtained from our algorithm

Prozac		Paxil		Seroquel		Olanzapine	
Variant	**Google Hits**	**Variant**	**Google Hits**	**Variant**	**Google Hits**	**Variant**	**Google Hits**
prozact	3960000	paxl	52300000	seroquels	1910000	olanzapin	1220000
prozaac	3160000	pxil	12200000	seroqul	1810000	olanzapoine	869000
prozaqc	1300000	pexil	10600000	seroqual	1810000	olanzapines	868000
prozaxc	1300000	paxol	2490000	sroquel	1800000	olanzaoine	864000
prozax	1270000	paxial	2340000	seruquel	1790000	olanzaopine	863000
prozc	1260000	paxiol	866000	saroquel	1760000	olanzapne	796000
prozec	1260000	paxill	856000	seroqel	1710000	olanzaplne	765000
proazac	1260000	paxilk	819000	seroquell	1230000	olanzapuine	734000
prozzac	1220000	paxilo	809000	serocquel	763000	olanzapins	567000
prazac	1210000	paxils	790000	seroguel	751000	olanzpine	565000
proazc	1180000	paxilv	750000	seroquol	742000	olanzopine	536000
proxac	1150000	paxilj	746000	sereoquel	676000	olanzipine	530000
prozacs	1120000	paxiln	738000	seriquel	615000	olanazapine	525000
prizac	1100000	paxilq	738000	serroquel	604000	olanzepine	386000
przac	1070000	paxcil	708000	serequel	111000	olanzapinm	6820
porzac	997000	paxiul	694000	seraquel	106000		
prozacc	995000	paxilz	668000	seroquela	5580		
prozaq	12500	paxila	5700				

tranSMART: An Open Source Knowledge Management and High Content Data Analytics Platform

Elisabeth Scheufele, MD, MS,[1,2] Dina Aronzon, MS,[1] Robert Coopersmith, Ph.D,[1] Michael T. McDuffie, MS,[1] Manish Kapoor, MS,[1] Christopher A. Uhrich,[1] Jean E. Avitabile,[1] Jinlei Liu, MS,[1] Dan Housman,[1] Matvey B. Palchuk, MD, MS[1,2]

[1] Recombinant By Deloitte, Newton MA; [2] Harvard Medical School, Boston, MA

Abstract

The tranSMART knowledge management and high-content analysis platform is a flexible software framework featuring novel research capabilities. It enables analysis of integrated data for the purposes of hypothesis generation, hypothesis validation, and cohort discovery in translational research. tranSMART bridges the prolific world of basic science and clinical practice data at the point of care by merging multiple types of data from disparate sources into a common environment. The application supports data harmonization and integration with analytical pipelines. The application code was released into the open source community in January 2012, with 32 instances in operation. tranSMART's extensible data model and corresponding data integration processes, rapid data analysis features, and open source nature make it an indispensable tool in translational or clinical research.

Background

Translation of biomedical research to patient care has been a difficult path, fraught with barriers and a paucity of information technology solutions. Funding structures have led to a landscape of siloed knowledge where collaboration and sharing were not facilitated.[1] Software tools and data handling processes have been lacking, with little success in standardizing data integration practices.[2] The pace of applying research findings into the clinical practice was also woefully slow.[3,4,5] These issues were rapidly gaining notice in the early 2000's, as identified in seminal publications.[6] In 2003, the National Institutes of Health defined the Roadmap to Medical Research, and identified translation as a "vital component of research and health-care improvement."[6] However, the journey to bring new findings to the point of care has been stymied at the clinical level.[3,4,6] Much of the research activity has resided in basic science and clinical trials (vs the clinical environment), where significant incentives exist to encourage development of new treatment options and modalities.

The demand for translating basic research into clinical practice has resonated in the pharmaceutical domain.[7] Johnson & Johnson (J&J) identified a "significant challenge in the lack of translatability of preclinical models into meaningful biological knowledge."[8] J&J partnered with Recombinant Data Corporation in 2008 to develop "a knowledge management platform that would provide access to all R&D data as well as advanced analytics."[8] The first iteration of tranSMART was deployed in 2010; it utilized components of the i2b2 (Informatics for Integrating Biology and the Bedside; http://www.i2b2.org) schema,[9] and incorporated search capabilities. A parallel goal was to release the software as open source to encourage precompetitive data sharing among pharmaceutical entities and academic institutions, including the Cancer Institute of New Jersey.[8] Since its open source release in January 2012, tranSMART has continued to mature into a robust translational research knowledge management and analysis platform.

Materials and Methods

Application Infrastructure

tranSMART version 1.1 employs an N-tier architecture (see Figure 1) and is built on Grails – a rapid web application development framework based on a Java platform (http://www.grails.org). The N-tier architecture typically separates application needs into a data-handling tier, a business logic tier, and the presentation tier. Here the data-handling tier is supported by a relational database with a series of purpose-built schemas to house various data types. The business logic tier code base is also written in Grails. The presentation layer is a web-based application. As a framework, Grails was selected for tranSMART because of its value in facilitating the development of robust web applications. By its nature, the framework lends itself to extensibility, making it a good fit for open source development. Functionality can be added to Grails applications through the plugin infrastructure, with plugins that are custom built or sourced through the Grails community.

The Data Tier (see Figure 1) houses the data used in the tranSMART in relational databases, including raw data from files and databases as well as data normalized to common formats and ontologies (discussed in the Data section below). Supported file types include Microsoft Word documents, text files, PDFs, spreadsheets, and 'omic data files. Data can be added to the tranSMART database through a suite of tools that support data extraction, transformation and loading (ETL). These tools load raw data that has been mapped via standard templates by a data curator into tranSMART schemas. Additional data files of any format can be associated on the patient level and linked via the file system. Grails makes use of a custom Object Relational Mapping system called GORM. GORM provides a connection between programmatic objects in Grails and underlying database objects. One significant advantage of the GORM layer is that it allows for the application to be database agnostic, thus enabling tranSMART to run on virtually any Relational Database Management System (e.g., Oracle, PostgreSQL). The Data Tier is made accessible through data services, which are predefined methods of communication between various components of the overall application. Data services reside in the business logic tier.

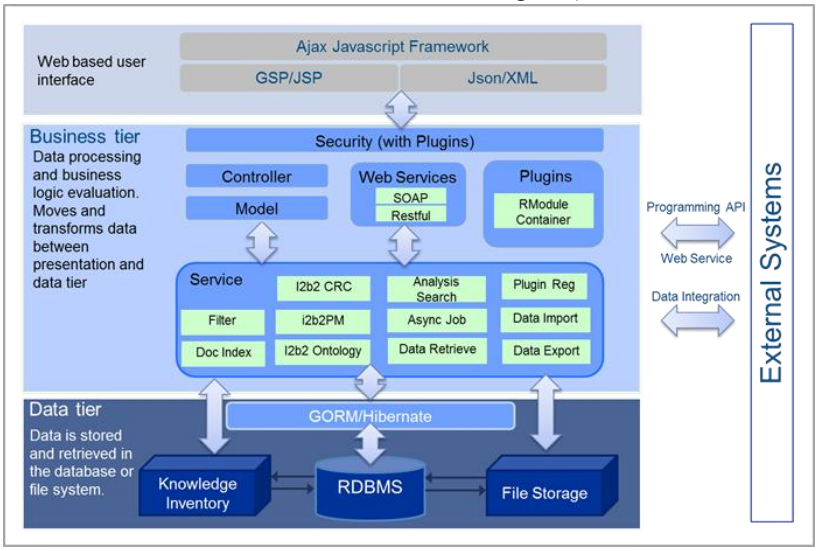


Figure 1. N-tier Architecture of tranSMART

The Business Logic Tier (see Figure 1) encapsulates tranSMART's application logic and provides services that implement tranSMART's core functionality such as security, i2b2 features, data export and other plugins. Security in tranSMART is handled by the Spring Security Core open source Grails plug in. Spring Security is the *de facto* authorization plugin adopted by the Grails community; it allows for the rapid development of form-based or single sign-on authentication. Built into Spring Security is a role-based permissions management system. Another service is the i2b2 interface module that provides connectivity to the various individual i2b2 applications (known as cells) through web services and direct database calls. There is an export functionality that allows the user to download datasets in commonly available formats such as tab delimited text files, GSEA, and PLINK. The exported data can be used for further analysis in external statistical packages (e.g., Stata) Additionally, an Rmodules Grails plugin allows a researcher to specify a cohort of subjects with associated data of interest and run pre-defined analytics via an R statistical package. To do so, a non-programmer user chooses one of the available analyses and specifies the relevant input parameters. The data is subsequently pulled, formatted, and sent via a tranSMART API to R, where the relevant scripts are executed. The results are then presented in the tranSMART interface. Together, the Business Tier services are responsible for the core functionality of tranSMART.

tranSMART's Presentation Tier (see Figure 1) relies on the data services to decouple the presentation layer from the data. JavaScript, along with popular libraries such as jQuery and ExtJS, is used to define the UI elements and the interactions among them. Asynchronous JavaScript and XML (AJAX) are used extensively to provide a rich interactive experience. JavaScript Object Notation (JSON) and XML provide a standard communication language between the presentation and business layers. Groovy Server Pages (GSP) are used by the Grails framework to display HTML content to the end user. Tag libraries can be built or downloaded to allow for more rapid development of common form and display elements.

Data, Data Store and Data Integration

The tranSMART knowledge management platform allows for the integration of data from a variety of data sources, across multiple data types. The data types that can be loaded into tranSMART include patient- or study-level clinical data (e.g., demographics, diagnosis, medications, lab results, etc.), subject-level high dimensional data (e.g., genotyping calls, gene expression arrays, protein expression arrays, etc.), study-level results and findings, as well as study descriptors in the form of metadata. The data can come from public sources (e.g., TCGA, GEO), or from internal sources (e.g., institutional clinical trials, etc.). Reference content brought into tranSMART can be

proprietary (e.g., GeneGo, Ingenuity) or openly available (e.g., Entrez, MeSH). Additionally, the original data sources can be stored as files and accessed for export via the user interface for future research workflows.

Data integration is a vital step in ensuring downstream success of any data analysis undertaking. There are multiple issues in the data management workflow in a typical research environment. Raw data frequently have formatting problems and form discrepancies that are undetected or unintended at the original capture level, but can cause problems in subsequent activities, such as search or data analysis. Data often require some level of normalization or harmonization in order to be analyzed and utilized properly. In the typical research workflow, data is frequently cleaned to a very specific analytical purpose. If another type of analysis is desired, the data requires another round of time consuming data cleaning activities. If data is instead conformed to a common structure and representation of meaning, thus rendering it agnostic to the analysis, users can focus on studying the data rather than spending valuable time on data preparation.

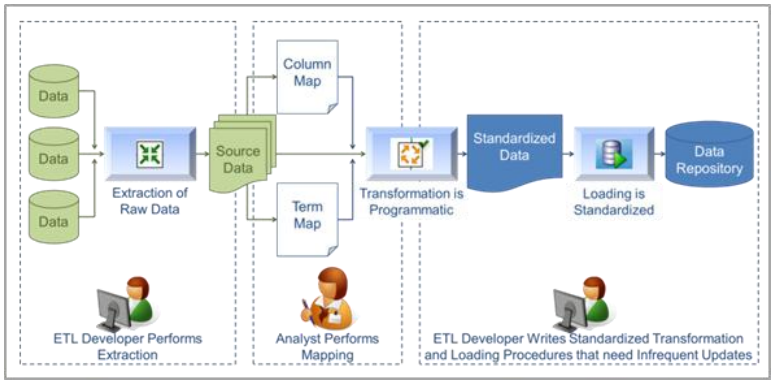


Figure 2. Best Practice Approach to Data Integration

We employ an internally developed best practices approach to data integration from various sources into a common environment (see Figure 2).[10] This robust process systematizes the data integration effort into a series of well-defined steps and delegates specific tasks to staff members with appropriate skillsets. Data is mapped to a standard ontology that can be institution-specific or conform to an industry standard (e.g., SNOMED CT, SDTM).

The data integration process requires several discrete steps. First, under the direction of the principle scientist or system product owner, the data source is identified and selected for migration into the tranSMART environment. Next, an ETL developer performs the source data extraction and the data curation process is initiated. Both syntactic and semantic mappings are created in this step (see Figure 2). A knowledgeable analyst who follows standard practices performs the mapping of the data into the templates, ensuring high quality output. Once the mapping templates are filled out, an ETL developer can run the programmatic transformation of the source data into the

Category	Type	Description	Example	Usage	Storage
Level 1	Raw	• Raw data from source platform • Not normalized	• Raw binary machine reads • Data on the Case Report Form	• Processing pipeline • Dataset Explorer export	File system
Level 2	Processed	• Normalized data through curation or data processing pipelines	• Clinical trial data • RMA or MAS5 normalized gene expression data • SNP data with calls and CNV	• Dataset Explorer	Database: DeApp, i2b2DemoData
Level 3	Interpreted	• Interpreted or aggregated data from processed data	• Z-scores for gene expression data • Survival times calculated at the end of a study	• Dataset Explorer • Search	Database: DeApp, BioMart
Level 4	Summary and Findings	• Quantified association and analysis across multiple samples. • Published results	• Fold changes • GWAS • Results from publications	• Search	Database: BioMart
Master Data	Slow changing data	• Data about key business entities in the system. • Data might be from internal or external data source.	• Study design • Platform specifications • User defined gene lists	• Dataset Explorer • Search	Database: i2b2Metadata, i2b2DemoData, BioMart, SearchApp
Reference Data	Slow changing data used as reference	• Data from other system that's used as identifier data or as a reference to other systems	• Affymetrix annotation files • Gene ID's from Entrez • Disease lists from MeSH	• Dataset Explorer • Search	Database: DeApp, BioMart
MetaData - Structural	Metadata	• Data that describes data structure	• Data dictionary • Schema guide	• Documentation	File system
MetaData – Administrative (Operational)	Metadata	• Data associated with application/data access and operation	• ETL auditing and QC results • Application access results	• Search	Database: searchApp, rdc_cz

Table 1. Data Categories Supported by tranSMART platform

standard format and load the output into tranSMART. In the last stage of the process, the data undergoes a quality assurance review to ensure integrity, and any defects in the data are noted and communicated.

The data model supporting tranSMART segregates the content into multiple stores and optimizes data structure to represent each type of content. For example, patient-centric clinical data is stored in the i2b2 star schema. High-dimensional data is housed in a different data store where each of the data types (e.g., SNP) retains its specific structure. There are many categories of data that are loaded into tranSMART (see Table 1). For levels 1-4, we employ definitions summarized by the Cancer Genome Atlas (http://www.cancergenome.nih.gov/), which indicate the degree of processing that the data has undergone. Master data are slowly changing elements such as patient identifiers, and Metadata are associated data of a structural or administrative nature. Another feature of the data model is that all the core data elements, such as genes, pathways, diseases, compounds, and concept codes, have a unique ID (UID). The crosswalks among different data stores are enabled using these UIDs. As data is reloaded or appended into the system, the relationships established with the resident data are maintained via the UIDs.

Search

The Search performs fast and comprehensive queries against the tranSMART repository of research and reference data. This feature relies on the Apache Lucene Solr (http://lucene.apache.org/solr/) search server – an open source, robust search engine that supports near-real time indexing capabilities. Solr indexes files and table records when the data is initially loaded into the tranSMART databases. tranSMART allows click-through to externally available resources, such as PubMed and Entrez. Users benefit from a richer experience since searches for diseases or genes yield the Entrez 'omic details or PubMed articles in addition to the data stored locally in tranSMART.

Dataset Explorer

Dataset Explorer is the principle access point for primary study data. Within Dataset Explorer, the user is presented with data in an organized hierarchical fashion. Using drag and drop, the user can execute a variety of analytic workflows (or pre-defined statistical algorithms). The user is able to specify a cohort of patients, choose an available analysis modality, and set the relevant parameters. Once the selections are made, the scripted pipelines are executed. The results are returned to the user and presented in a graphical fashion that best matches the completed analysis. The user may continue to explore the data further by altering cohort details or analytic parameters and running other pre-scripted analyses. Thus the user is able to generate a hypothesis within tranSMART using one dataset and then test this hypothesis on a different dataset. Generated results can be saved or exported. No knowledge of statistical scripting languages is necessary to analyze the data because the user is presented with a predefined menu of powerful analysis options.

tranSMART performs its default analyses without allowing end-users direct access to raw data, thus preserving the integrity of the underlying sources. For researchers with deeper statistical programming knowledge, tranSMART allows direct manipulation of the data via export or direct connection to R. The latter uses a set of pipelines from the tranSMART database to the R statistical package and a custom command library to support direct interaction with the data in R.

Security

The tranSMART application uses a role-based security model to enable the use of the platform across large organizations. User authentication can be integrated into an organization's existing infrastructure to streamline user management. The security model allows an organization to control data access in accordance with internal policies governing the use of research data. Security is based on the user's role, where defined roles have different levels of data access. Many implementers have integrated security with proprietary single sign-on solutions, however there is also interest in utilizing open source options, such as Shibboleth (http://shibboleth.net/).

Results

The tranSMART application code became available as open source in January 2012 and is licensed through GPL 3. Currently, there are 32 implementations of tranSMART (see Table 2) in operation, with 11 implementations with pharmaceutical companies, 6 with research institutions, and 5 with non-profit organizations. Also, the tranSMART community has developed a strong web presence. The website, http://transmartproject.org/, is the hub of the open source community. The site acts as a portal for the central repository of code (both core framework and contributions), as well as the wiki with project information and active discussion groups. There is a significant social media presence with handles on LinkedIn, Twitter (@transmartapp) and Google Groups. The community is thriving with active engagement, collaboration and contributions.

University of Michigan, Pistoia Alliance and Imperial College of London initiated the tranSMART Foundation (http://www.transmartfoundation.org) in 2013. The goal of the tranSMART Foundation is to coordinate development of the next generation of tranSMART as it evolves into a global platform solution for clinical and translational research.

There have already been significant contributions to the tranSMART application programming interface (API) since its release. Dataset Explorer has been improved with the addition of a data export feature, as well as a number of new statistical analyses in the Advanced Workflows. The R advanced analysis functions were contributed by the open source community and include: box plot with ANOVA, scatter plot with linear regression, table with Fisher test, survival analysis, heat map, hierarchical clustering, k-means clustering, marker selection, principle component analysis (PCA), correlation analysis, and line graph. The Search feature has been expanded with the incorporation of a faceted search capability as well as the extension of search to support Genome Wide Association Studies. The open source community is demonstrating excitement about the potential of tranSMART and is actively working to expand on the current platform with additional features in the development pipeline.

Organization Types	# Instances
AMC	4
Biopharma	11
Cancer Center	2
Commercial Software	1
Government	4
Non-profit	5
Research	6

Table 2. tranSMART Implementations

Discussion

tranSMART is an open source knowledge management and high content analytics platform that enables secondary use of clinical and translation research data. Typically, researchers do not have unfettered access to the necessary data because of poor availability. In addition, raw data requires significant curation to prepare it for analysis. tranSMART addresses data availability by acting as a catalog of research data and related knowledge such as reference data (e.g., probe to gene mapping, study metadata, links to publication repositories and 3rd party references, etc.). tranSMART enables rapid exploration of research trials data in an ad hoc fashion by allowing users to identify cohorts, and perform pre-defined univariate statistical analyses with the ease of a drag and drop interface and without prior statistical programming knowledge. tranSMART supports data sharing and promotes collaboration at the institution level by providing an environment where individual departments can access research data from the entire institution.

A major component to the success of tranSMART is the employment of a common data model. By defining a common structure with common meaning for data, the platform realizes upstream benefits by being able to scale data curation activities according to breadth and depth of source data available, as well as downstream benefits by reusing processing pipelines and analytic tools for optimal and efficient end user experience.

Success of an open source project depends largely on the community engagement. Necessary technological underpinnings for a healthy community have been put in place: robust web presence, development tools (code repository, issue tracker, developer forum) to support tranSMART in this purpose. A flourishing community, as exampled by the bevy of contributions to the source code, is the most telling feature of an open source project; tranSMART appears to be well on its way to achieving this goal.

The tranSMART application platform is enjoying a wave of rapid adoption across life sciences and health care provider sectors. About a year after release to open source, 32 institutions have implemented the platform for use in their research and development activities. Individual implementations range from proof of concept to full institutional implementations. The platform is in the early stages of adoption, but there are already publications demonstrating the use of tranSMART.[11] The highest adoption is seen in the pharmaceutical industry, with noteworthy representation by government agencies, academic medical centers and research institutions. The tranSMART Foundation is collaborating internationally to support institutions outside the US in their use of tranSMART.

A combination of tranSMART's architecture and associated data integration processes serve as the foundation for a robust translational research environment. tranSMART supports multiple data types of varying levels of complexity and relationships, and provides novel yet intuitive end-user components for accessing and analyzing large volumes of data quickly in a secure environment. These features bring together complex patient phenotypic and high dimensional 'omics data, collected for both research and in EHRs, and enable the continuum of translational science from basic research to bedside in a single application platform.

References

1. Zerhouni EA. Translational and clinical science–time for a new vision. *N Engl J Med*. 2005;353(15):1621-1623.
2. Sung NS, Crowley WF Jr, Genel M, Salber P, et. al. Central challenges facing the national clinical research enterprise. *JAMA*. 2003;289(10):1278-1287.
3. Contopoulos-Ioannidis DG, Alexiou GA, Gouvias TC, et. al. Life cycle of translational research for medical interventions. *Science*. 2008;321:1298-1299.
4. Ioannidis JPA. Commentary: materializing research promises: opportunities, priorities and conflicts in translational medicine. *J Transl Med*. 2004;2:5:1-6.
5. Lenfant C. Clinical research to clinical practice–lost in translation? *N Engl J Med*. 2003;349:868-874.
6. Drolet BC, Lorenzi NM. Translational research: Understanding the Continuum from Bench to Bedside. *Transl Res*. 2011;157:1-5.
7. Szalma S, Koka V, Khasanova T, et. al. Effective knowledge management in translational medicine. *J Transl Med*. 2010;8:68:1-9.
8. Perakslis ED, Van Dam J, Szalma S. How informatics can potentiate precompetitive open-source collaboration to jump-start drug discovery and development. *Clin Pharmacol and Ther*. 2010;1-3.
9. Murphy SN, Weber G, Mendis M, et. al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17:124-130.
10. Aronzon D, Palchuk MB. Best practices in biomedical data extraction, transformation and load. *AMIA TBI Summit*. 2013.
11. Irgon, J, Huang CC, Zhang Y, Talantov D, Bhanot G, Szalma S. Robust multi-tissue gene panel for cancer detection. *BMC Cancer*. 2010;10:319.

Natural Language Processing Methods for Enhancing Geographic Metadata for Phylogeography of Zoonotic Viruses

Tasnia Tahsin[1], Rachel Beard[1], Robert Rivera[1], Rob Lauder[1], Garrick Wallstrom, PhD[1], Matthew Scotch, PhD, MPH[1], Graciela Gonzalez, PhD[1]
[1]Arizona State University, Tempe, AZ, USA

Abstract

Zoonotic viruses represent emerging or re-emerging pathogens that pose significant public health threats throughout the world. It is therefore crucial to advance current surveillance mechanisms for these viruses through outlets such as phylogeography. Despite the abundance of zoonotic viral sequence data in publicly available databases such as GenBank, phylogeographic analysis of these viruses is often limited by the lack of adequate geographic metadata. However, many GenBank records include references to articles with more detailed information and automated systems may help extract this information efficiently and effectively. In this paper, we describe our efforts to determine the proportion of GenBank records with "insufficient" geographic metadata for seven well-studied viruses. We also evaluate the performance of four different Named Entity Recognition (NER) systems for automatically extracting related entities using a manually created gold-standard.

Introduction

Zoonotic viruses, viruses that are transmittable between animals and humans, have become increasingly prevalent in the last century leading to the rise and re-emergence of a variety of diseases[1]. In order to enhance currently available surveillance systems for these viruses, a better understanding of their origins and transmission patterns is required. This need has led to a greater amount of research in the field of phylogeography, the study of geographical lineages of species[2]. Population health agencies frequently apply phylogeographic techniques to trace the evolutionary changes within viral lineages that affect their diffusion and transmission among animal and human hosts[3, 4, 5].

Phylogeography depends on the utilization of both sequence and location data which are often obtained from online resources such as GenBank (http://www.ncbi.nlm.nih.gov/genbank/) .While there is an abundance of sequence data records in GenBank, many of them lack sufficient geographical metadata that would enable specific identification of the isolate's location of collection. In our previous study[6] we found that the geographic information of 80% of the GenBank records associated with single or double stranded RNA viruses within tetrapod hosts was less specific than 1st level administrative boundaries (ADM1) such as state or province. More detailed information concerning the location from which sequences were collected is often available within their corresponding journal articles. However, this manual process does not allow researchers to conveniently retrieve the needed data. We investigated the potential of natural language processing (NLP) to enhance the geographical data available for phylogeography studies by extracting information on the origin of viruses from natural language text[6]. In order to accurately link each GenBank record to its corresponding location of isolation, NER systems can be used to identify additional entities other than location within the text. Such entities can be used to provide extra information that the system can use to distinguish the correct location from other location mentions. Previously we used BANNER and the Stanford NER tool to automatically tag gene and location mentions in text respectively. Here we build on this work by proposing an automated process through which spatial information, dates, genes and species mentions are extracted from journal articles. We validate these results by using a gold standard developed through manual annotation.

Background

Phylogeography has the potential to inform population health endeavors by studying the spread of viruses in relation to migration and host populations. For instance, Gray et al. explores the spatial distribution of Rift Valley Fever Virus (RVFV) within Africa using information regarding the time and location of sample collection combined with sequence data[4]. Using the work of Lemey et al. 2009 as a guide, phylogeographic analysis was performed using the BEAST software package in order to reconstruct viral transmission and relate these findings to data for livestock population density[7, 8]. This work demonstrates the benefits of phylogeographic study by highlighting the need to reassess assumptions regarding the spread of RVFV via sheep and cattle populations, considering the spread of the virus over large low density area[4]. In another example, Weidmann et al. studied the expansion of tick-borne encephalitis virus (TBEV) within central Europe, as opposed to the observed spread throughout the Eurasian region

by examining E gene sequences[5]. Following Bayesian analysis, it was found that west to east viral migration was indicated for central Europe as opposed to an east to west direction in Eurasia. In addition, it was concluded that the emergence of multiple subclades could be explained by recent evolutionary bottlenecks[5].

Previously, there has not been work aimed at the automated extraction of spatial data for the enhancement of phylogeography. However, other applications such as geographical information retrieval (GIR) share similar challenges in processing natural language documents containing geographic mentions despite having different goals. Bordogna et al. addressed this subject by developing a GIR model that represents the uncertainty of geographic context of texts by associating fuzzy descriptors of specific locations which indicate the author's perception[9]. Other approaches to bio-surveillance have utilized NLP techniques such as NER and n-grams to scan and classify online articles concerning disease outbreaks of potential interest to public health[10,11]. While these works share similar endeavors, we address the complex issue of improving geospatial data availability relating to genetic sequence data. Furthermore, the framework we develop here will be applied towards the extraction of geospatial data of interest which will allow the linkage of specific mentions of a GenBank record and a location, thereby aiding the development of future phylogeographic models.

Materials and Methods

The process undertaken to complete this study can be divided into five distinct stages: selection of the viruses, extraction of relevant GenBank data related to each virus, computation of "sufficiency" statistics on the extracted data, development of an integrated NER system, and evaluation of this system on a manually annotated corpus of full-text PubMed articles. Figure 1 in Appendix A provides a brief overview of each of these steps. A detailed description of each phase is given below.

Virus Selection and GenBank Data Extraction: The domain of this study was limited to zoonotic viruses that are most consistently documented and tracked by public health, agriculture and wildlife state departments within the United States. These viruses include influenza, rabies, hantavirus, western equine encephalitis (WEE), eastern equine encephalitis (EEE), St. Louis encephalitis (SLE), and West Nile virus (WNV). The Entrez Programming Utilities (E-Utilities) was used to download the following fields from 59,595 GenBank records associated with these viruses: GenBank Accession ID, Pubmed Central ID, Strain name, Collection date and Country. This set of records consisted of all records in GenBank related to the selected viruses that contained Pubmed Central (PMC) IDs for referencing articles. We limited our set to records with PMC IDs since the NER system being tested is only relevant for these records. Figure 2 in Appendix A provides a screenshot of the extracted data.

Sufficiency vs. Insufficiency Analysis: The data extracted from Genbank was used to compute the percentage of GenBank records that had insufficient geographic information for each of the selected viruses. In order to perform this computation, we used data from the ISO 3166-1 alpha-2 table[12] and the GeoNames[13] database. The ISO 3166-1 alpha-2 is the International Standard for representing country names using two-letter codes. The GeoNames database contains a variety of geospatial data for over 10 million locations on earth, including the ISO 3166-1 alpha-2 code for the country of each location and a feature code that can be used to determine the administrative level of each location. To allow for efficient querying, we downloaded the main GeoNames table and the ISO alpha-2 country codes table from their respective websites and stored them in a local SQL database. Prior to adding the ISO data to the database, some commonly used country names and their corresponding country codes were added to the table since it only included a single title for each country. For example, the ISO table included the country name "United States" but not alternate names such as "USA", "United States of America", or "US". Using the created database in conjunction with a parser written in Java, we were able to retrieve most of the geographic information present within the records and classify each of them as sufficient or insufficient.

For the purpose of this project, we considered any geographical boundary more specific than ADM1 to be "sufficient". Based on this criterion, a feature code in GeoNames was categorized as sufficient only if it was absent from the following list of feature codes: ADM1, ADM1H, ADMD, ADMDH, PCL, PCLD, PCLF, PCLH, PCLI and PCLS. The method for evaluating the geographical sufficiency of a GenBank record was dependent upon whether the record included a country name. A GenBank record with a country mention was called sufficient if the geographic information extracted from that record included another place mention whose feature code fell within the class of sufficient feature codes and whose ISO country code matched that of the retrieved country. Place mentions with matching country codes often had several different feature codes in GeoNames. Such places were only called

sufficient if all feature codes corresponding to the given pair of place name and country code were classified as sufficient. In cases where the GenBank record had no country mention, the record was called sufficient only if all matching GeoNames entries for any of the places mentioned in it had sufficient feature codes. The sufficiency criteria were designed to ensure that a geographic location is only called sufficient if its administrative level was found to be more specific than ADM1 without any form of ambiguity. Figure 3a in Appendix A illustrates the pathways of geographical sufficiency for GenBank records in a diagram.

In order to obtain the geographic information for each Genbank record, we used a Java parser which automatically extracted data from the "country" field of each record. Since the "country" field typically contained multiple place mentions divided by a set of delimiters consisting of comma, colon and hyphen, we first split this field using these delimiters. We then checked each string obtained through this process against the ISO country code table to determine whether it was a potential country name for the record's location. If the query returned no results, then the locally stored GeoNames tables was searched and for each match found, the corresponding ISO country code and feature code were extracted. Figure 3b in Appendix A on the right shows a diagram of this process.

In cases where no sufficient location data was found from the "country" field of a GenBank record, the Java parser searched through its "strain" field. This was done because some viral strains such as influenza include their location of origin integrated into their names. For example, the influenza strain "A/duck/Alberta/35/76" indicates that the geographic origin of the strain is Alberta. The different sections of a strain field are separated by either forward slash, parenthesis, comma, colon, hyphen or underscore and so we used a set of delimiters consisting of these characters to split this field. Each string thus retrieved was queried as before on the ISO country code table and the GeoNames table. GeoNames often returned matches for strings like 'raccoons' and 'chicken' which were actually meant to be names of host species within the "strain" field and so a list of some of the most frequently seen host name mentions in these records was manually created and filtered out before querying GeoNames.

Some of the place mentions contained very specific location information which resulted in GeoNames not finding a match for them. A list was created for strings like 'north', 'south-east', 'governorate' etc. which when removed from a place mention may produce a match. In cases of potential place mentions which contained any one of these strings and for which GeoNames returned no matching result, a second query was performed after removal of the string.

Development and Evaluation of Integrated NER System: An NER system for identifying species, gene, date and location mentions in text was developed by integrating LINNEAUS[14], BANNER[15], GeoNamer and Stanford SUTime[16]. LINNEAUS, BANNER and Stanford SUTime are widely-used, state-of-the-art open source NER systems for recognition of species, gene and temporal expressions respectively. A detailed description of each of these tools is provided in the Appendix. GeoNamer is a dictionary-based location tagging system that we built using GeoNames. The dictionary used by GeoNamer was created by retrieving distinct place names from the GeoNames table and filtering out commonly used words from the retrieved set. Words filtered out include stop words, generic place names such as 'cave' and 'hill', numbers like 'one', domain specific words such as 'biology' and 'DNA', most commonly used surnames like 'Garcia', commonly used animal names such as 'chicken' and 'fox' and other miscellaneous words such as 'central'. This was a crucial step since the GeoNames databases contains a wide array of commonly used English used words which may cause a large volume of false positives if not removed. The final dictionary consisted of 5,396,503 entries. To recognize place mentions in a given set of text files, GeoNamer first builds a Lucene index on the contents of the files. It then constructs a phrase query for every entry in the Geonames dictionary and runs each query on the Lucene index. The document id, query text, start offset and end offset for every match found is written to an output file.

The developed system was tested on a set of twenty-seven manually-annotated full-text PubMed Central articles downloaded manually in the pdf version and converted to text using Adobe Acrobat. The number of papers selected for each virus was influenced by the number of GenBank records with PMC ids that were available for the given virus. Ten papers were selected for rabies, nine for influenza, two for hantavirus, WEE and WNV and one for SLE and EEE. The articles for each virus were chosen by using Excel's RAND function on the subset of extracted GenBank records related to that virus which had insufficient geographic information.

Three annotators tagged the following five entities in each article using the freely available annotation tool, BRAT[17]: gene names, locations, dates, organisms and viruses. A detailed description of the annotation guidelines is given in the appendix. Before creating the guidelines, each annotator individually annotated six common articles and

compared and discussed their results to devise a reasonable set of rules for annotating each entity. After discussion, the annotators re-annotated the common articles based on the guidelines and divided the remaining articles amongst themselves. The inter-annotator agreement was calculated for each pair of annotators and the entity taggers were evaluated on the annotated corpus.

Results

Sufficiency vs. Insufficiency Analysis: The results of the sufficiency vs. insufficiency analysis are given in Table 1. 64% of all GenBank records extracted for this project contained insufficient geographic information. Amongst the seven studied viruses, WEE had the highest and EEE had the lowest percentage of insufficient records.

Table 1. Percentage of GenBank records with insufficient geographic information for each virus.

Submission Type	Number of Entries	% Insufficient
WEE	67	90
Rabies	4450	85
WNV	1084	79
SLE	141	74
Hanta	1745	66
Influenza	51734	62
EEE	374	51
All	59595	64

Inter-rater Agreement: The results for the comparison of the annotations performed by our three annotators on 6 common papers can be found in Table 3 of Appendix B. We used both the Jaccard Similarity and the harmonic mean of the intersection between two annotators divided by the total number annotated by each as a measure of inter-rater agreement and had over 90% agreement with overlap matching and over 86% agreement with exact matching in all cases.

Performance Analysis of NER Systems: The performance metrics for the NER systems at tagging the desired entities in the test set are listed in Table 2. The highest performance was achieved by Stanford SUTime for date tagging. Tagging of genes had the lowest performance.

Table 2. Performance statistics of the integrated NER system

Entity	Precision (Exact;Overlap)	Recall (Exact;Overlap)	F-measure (Exact;Overlap)
GeneName	0.070; 0.239	0.114; 0.395	0.087; 0.297
Location	0.452; 0.626	0.658; 0.783	0.536; 0.696
Species	0.853; 0.962	0.563; 0.658	0.678; 0.781
Date	0.800; 0853	0.681; 0.727	0.736; 0.785

Discussion

Based on our analysis, at least half of the GenBank records for each virus lack sufficient geographic information and the proportion of insufficient records can be as high as 90%. The virus with the highest level of sufficiency, WEE, had a large number of records with county level information in the "country" field. However, the insufficient records for this virus typically contained no place mention, not even at the country level. A key reason for our calculated percentage of sufficient GenBank records being higher for these seven viruses than what we previously computed in Scotch et al.[6] was the inclusion of the "strain" field. The "strain" field often contained specific location information which, when combined with place mentions present within the "country" field, made the record geographically sufficient. The virus for which the inclusion of "strain" field had the greatest impact on boosting the sufficiency percentage was influenza. Most of the GenBank records associated with this virus had structured "strain" fields from which the parser could easily separate place mentions using GeoNames.

Although the sufficiency classifications produced by our system were correct most of the time, there were a few cases where a record got incorrectly labeled as insufficient even when it contained detailed geographic information. This typically happened because GeoNames failed to return matching results for these places. For instance, the country field "India: Majiara,WB" was not found to be sufficient even though Majiara is a city in India because GeoNames has no entry for it. In some cases the lack of matching result was due to spelling variations of the place name. For instance the country field "Indonesia: Yogjakarta" was called insufficient since "Yogjakarta" is spelled as "Yogyakarta" in GeoNames. Sometimes the database simply did not contain the exact string present in the GenBank record. For instance, it does not have any entry for the place "south Kalimantan" but it contains the place name "kalimantan". Errors due to inexact matching were greatly mitigated by removing strings such as "south" from the place mention, as described in the "Methods" section.

Most of the NER systems performed significantly better with overlap measures than with exact-match measures. This is because our annotation guidelines typically involved tagging the longest possible match for each entity and the automated systems frequently missed portions of each annotation. Stanford SUTime had the best overlap f-measure of 0.785, closely followed by LINNEAUS with an overlap f-measure of 0.781. Although Stanford SUTime was fairly effective at finding date mentions in text, it tagged all four-digit-numbers such as "1012" and "2339" as years, leading to a number of false positives. The poor recall of LINNEAUS was mostly caused because the dictionary used by LINNEAUS tagged only species mentions in text while we tagged genus and family mentions as well. It also missed a lot of commonly used animal names such as monkey, bat, badger and wolf. GeoNamer was the third best performer with the highest recall but second lowest precision. This is because the GeoNames dictionary contains an extensively large list of location names, many of which are commonly used words such as "central". Even though we filtered out a vast majority of these words, it still produced false positives such as "wizard". However, its performance was considerably better than that of the Stanford location tagger we used in our last paper. BANNER performed the worst amongst all the entity taggers. This is primarily because of the differences between the data set used to train the BANNER model and the annotation corpus used to test our system. The journal articles we selected had a large number of tables and BANNER was not able to identify the gene mentions in them. Instead, it tagged several entries within the table as a single gene name. It also incorrectly tagged strings in all capital letters such as VEEV and H1N1 as gene names. As a result, BANNER had both poor recall and precision.

Although this study explores the problem of insufficient geographic information in GenBank more thoroughly than our prior paper[6], expanding the number of viruses that we included in our survey and loosening the definition of sufficiency to account for additional metadata present in the "strain" field, the number of papers annotated as the gold standard increased, but is still limited. Thus, the performance of the taggers reported can be construed as a preliminary estimate at best. The set of taggers and their performance seem to be adequate for a large-scale application, with the exception of the gene tagger. However, we did not make any changes to the BANNER system (specifically, re-training). We had used it before and we had already identified limitations, but changes to it are not possible until sufficient data is annotated for retraining.

Conclusion

It can be concluded that the majority of GenBank records for zoonotic viruses do not contain sufficient geographic information concerning their origin. In order to enable phylogeographic analysis of these viruses and thereby monitor their spread, it is essential to develop an efficient mechanism for extracting this information from published articles. Automated NER systems may help accelerate this process significantly. Our results indicate that the NER systems LINNEAUS, Stanford SUTime and GeoNamer produce satisfactory performance in this domain and thus can be used in the future for linking GenBank records with their corresponding geographic information. However, the current version of BANNER is not well-suited for this task. We will need to train BANNER specifically for this purpose before incorporating it within our system.

Acknowledgments

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number R56AI102559 to MS and GG. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

References

1. Krauss, H. (2003). Zoonoses : infectious diseases transmissible from animals to humans (3rd ed.). Washington, D.C.: ASM Press.
2. Avise, John C. (2000). Phylogeography : the history and formation of species Cambridge, Mass.: Harvard University Press.
3. Ciccozzi M, et al. Epidemiological history and phylogeography of West Nile virus lineage 2. Infection, Genetics and Evolution. 2013:17;46-50.
4. Gray RR, and Salemi M. Integrative molecular phylogeography in the context of infectious diseases on the human-animal interface. Parasitology-Cambridge. 2012;139:1939-1951.
5. Weidmann M, et al. Molecular phylogeography of tick-borne encephalitis virus in Central Europe. Journal of General Virology. 2013;94:2129-2139.
6. Scotch, Matthew, et al. Enhancing phylogeography by improving geographical information from GenBank. Journal of biomedical informatics. 2011;44:S44-S47.
7. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol. 2009;5(9):e1000520.
8. Drummond AJ, et al. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. Aug. 2012;29(8):1969-1973.
9. Bordogna G, Ghisalberti G, and Psaila G. Geographic information retrieval: Modeling uncertainty of user's context. Fuzzy Sets and Systems. 2012:196;105-124.
10. Conway M, Doan S, Kawazoe A, and Collier N. Classifying disease outbreak reports using n-grams and semantic features. International journal of medical informatics. 2009;78:e47-e58.
11. Doan S, Vinh NTN, and Phuong TM. Classifying Vietnamese disease outbreak reports with important sentences and rich features. Proceedings of the Third Symposium on Information and Communication Technology. Aug. 23-24, 2012:260-265.
12. Iso.org. [Internet]. Genève. c2013. [cited 2013 Oct 10] Available from http://www.iso.org/iso/home/standards/country_codes.htm
13. Geonames.org. [Internet]. Egypt. c2013. [updated 2013 Apr 30; cited 2013 Sep 26] Available from http://www.geonamesorg/EG/administrative-division-egypt.html
14. Gerner M, Nenadic G, and Bergman CM. LINNAEUS: A species name identification system for biomedical literature. BMC Bioinformatics. 2010;11(85).
15. Leaman R and Gonzalez G. BANNER: An executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing. 2008;13:652-663.
16. Chang AX and Manning CD. SUTime: A Library for Recognizing and Normalizing Time Expressions.
17. Stenetorp P, et al. BRAT: A Web-based Tool for NLP-Assisted Text Annotation. EACL '12 Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012:102-107.

Appendix A

Figures

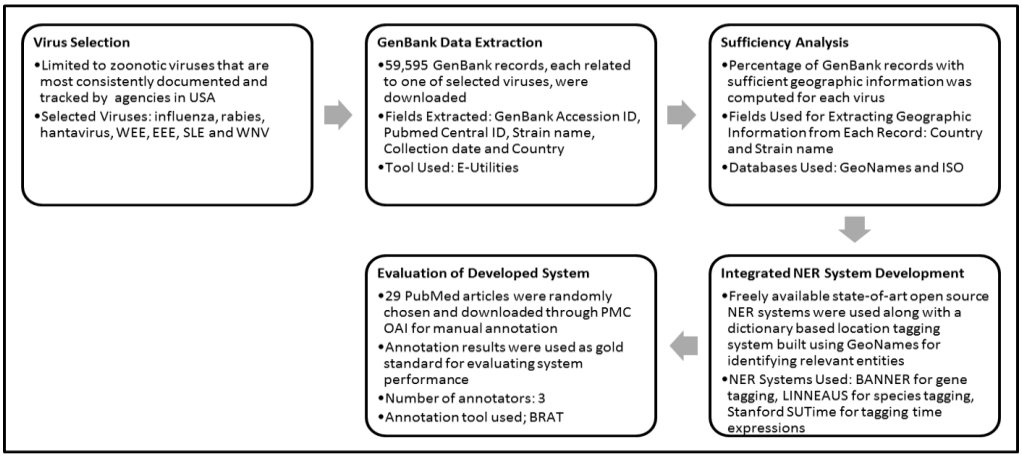


Figure 1. Flowchart of experiment procedure

	A	B	C	D	E
1	Accession ID	PMC ID	Strain_Name	Collection_Date	Country
2	JX912288	23951116	A/mallard/Sweden/50908/2006	08-Oct-200	Sweden: Ottenby
3	KF142499	23929468	A/swine/Korea/CY0423-12/2013	23-Apr-13	South Korea
4	CY146904	23908286	A/northern shoveler/California/2696/2011	29-Oct-11	USA: Solano County, CA
5	KF013908	23868121	A/sparrow/Guangxi/GXs-1/2012	10-Mar-12	China

Figure 2. Screenshot of data automatically extracted from GenBank

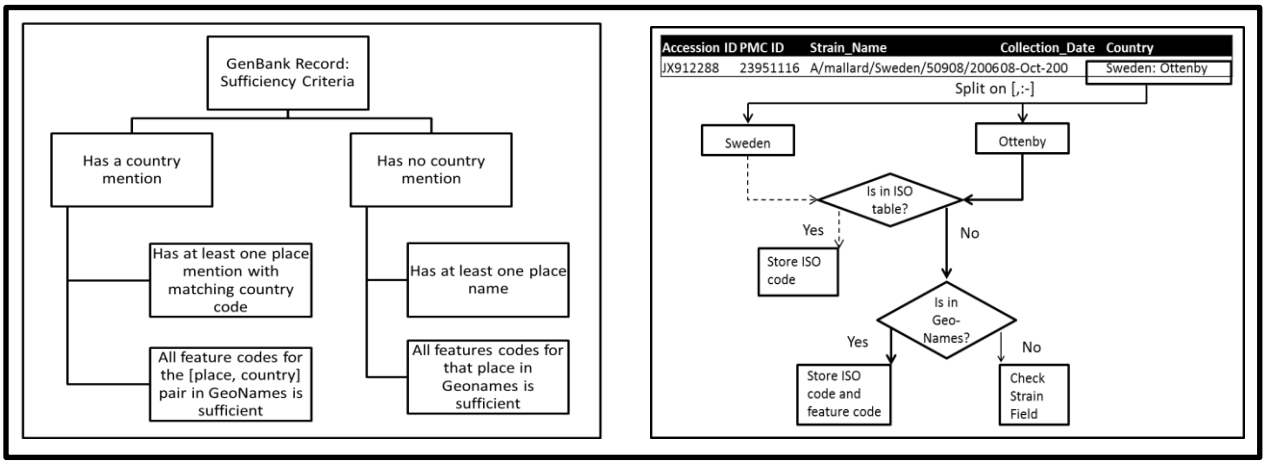


Figure 3. Description of Sufficiency criteria for GenBank record

Appendix B

Annotation Data

Table 1. Entity type frequency table

Entity	Annotator A	Annotator B	Annotator C
Date	386	387	390
GeneName	230	209	208
Location	846	846	903
Organism	916	866	850
Virus	1037	994	1031

Table 2. Additional inter-rater agreement measurements

Entity	$\frac{A \cap B}{A}$ (Exact;Overlap)	$\frac{A \cap B}{B}$ (Exact;Overlap)	$\frac{A \cap C}{A}$ (Exact;Overlap)	$\frac{A \cap C}{C}$ (Exact;Overlap)	$\frac{B \cap C}{B}$ (Exact;Overlap)	$\frac{B \cap C}{C}$ (Exact;Overlap)
Date	.977; .979	.974; .977	.984; .992	.974; .982	.966; .977	.959; .969
GeneName	.870; .880	.962; .976	.870; .887	.962; .981	.909; .952	.913; .957
Location	.943; .961	.948; .961	.939; .962	.879; .901	.944; .966	.885; .905
Organism	.885; .930	.935; .984	.843; .906	.902; .976	.906; .950	.924; .968
Virus	.932; .938	.972; .979	.945; .963	.951; .969	.965; .973	.930; .938

Table 3. Inter-rater agreement measurements (H.M. = Harmonic Mean, J.S. = Jaccard Similarity).

Entity	H.M.($\frac{A \cap B}{A}$, $\frac{A \cap B}{B}$) (Exact;Overlap)	H.M. ($\frac{A \cap C}{A}$, $\frac{A \cap C}{C}$) (Exact;Overlap)	H.M. ($\frac{B \cap C}{B}$, $\frac{B \cap C}{C}$) (Exact;Overlap)	J.S. (A,B) (Exact;Overlap)	J.S. (A,C) (Exact;Overlap)	J.S.(B, C) (Exact;Overlap)
Date	.975; .978	.979; .987	.962; .973	.952; .957	.950; .965	.928; .947
GeneName	.914; .926	.913; .932	.911; .954	.845; .868	.840; .872	.837; .913
Location	.945; .961	.907; .931	.914; .935	.897; .925	.831; .871	.841; .877
Organism	.909; .956	.874; .940	.915; .959	.833; .916	.792; .905	.843; .922
Virus	.952; .958	.947; .966	.947; .955	.907; .920	.903; .937	.900; .914
Mean	**.939; .956**	**.924; .951**	**.930; .955**	**.887; .917**	**.863; .910**	**.870; .914**

Appendix C

Annotation Schema

Entity Labels: Date, Location, Organism, Virus, GeneName

Entity Description and Tagging Guideline:

General

Non-entity specific guidelines for tagging: Instructions for tagging: When we have a group of words that collectively can be tagged as one entity, but individually mean another, then tag the group and the individual words separately. For example, “Eastern equine encephalitis” should be tagged as a virus, and separately “equine” should be tagged as an organism.

The components of a multi-word entity should not be tagged if the resulting tags represent the multi-word entity minus descriptive words. For example “fruit bat” should be tagged just as “fruit bat” rather than “fruit bat” and “bat” and “South Africa” should only be tagged “South Africa” rather than “South Africa” and “Africa”.

Strings that may represent multiple entity types should be tagged as all representative entities if the meaning is unclear in the eyes of the annotator. For example the string “Fort Morgan virus” would be tagged as “Fort Morgan” and “Fort Morgan Virus” and the entity types would be Location and Virus respectively. However, if the text were to

simply mention “Fort Morgan” without being followed by the string “virus” it cannot be assumed that this then is a Location entity. If the context does not make it abundantly clear which entity type it represents, then it would be tagged as all reasonable entity types.

Lastly, entity types should be identified based on their meaning within the text. For example, “tree” in “phylogenetic tree” would not be tagged as an organism and “Monte Carlo” in “Markov Chain Monte Carlo” would not be tagged as a location.

Date: Any date that identifies a decade or any time quantification more specific than that. Examples include “1970s”; “Jan 2013”; “June 16th, 1973”.

Instructions for tagging: For a specific date mention, group every adjacent component of the date into one tag. For example, “Jun 16th, 1973” should NOT be tagged in components “Jun”, “16th”, and “1973”. One tag should cover the entire date.

Special cases: In the cases of 10+ years ago, if a term of uncertainty is included before it (eg roughly, about) then do not tag it. Otherwise include “#### years ago” as a tag. If the reference to a date includes confidence intervals relating to the origin of speciation, for example “1,000 to 1,500 years ago,” then related dates should not be tagged.

If a range of dates is given, such as “1942 to 1952”, they should be tagged as separate entities. In the case of “May/June 1986,” this would be tagged as one date, otherwise information would be lost.

Location: Any named geographic location. Continents, countries, states, provinces, regions, territories, counties, named lakes, named mountain ranges, named deserts, named bodies of water, etc. General terms such as “the river”, “swamplands”, “in mountains” that cannot be used to identify specific locations should not be tagged.

Instructions for tagging: Include all parts of a location (as long as it provides more information) in the tag. For example, “upstate New York” and “South Lebanon” should include “upstate” and “South” in their tags.

Special cases: In cases like “Central and South America” tag the whole string as one entity, unless they could still be reasonably identified separately (Central has no specific meaning without America). Cases where a city and state are listed as [City, State] (or similar case), tag the city and state separately. Zip codes should be tagged as separate locations.

Organism: Any specific non-viral organism mentioned in the text. Broad terms such as “animals” should not be tagged.

Instructions for tagging: Include descriptive words attached to organism mentions if the descriptive word adds value beyond that which is implied. Such words may describe the region it's from, if it's domestic, or other types. Examples include “Canadian duck”, or “domestic poultry”. Examples of words that would not add value include “wild lion”.

Genus-species names should also be tagged (separate from its generic name). Abbreviations for organisms such as “gp” for guinea pig should be tagged as separate entities. If there is a mention of a genus or family name in text, this should be tagged as an organism separately from the common name of the organism.

Special cases: Different words addressing humans should be tagged: soldier, girl, boy, he, she, etc. Examples of words we would not tag are “trees” in reference to phylogenetic trees, or “host” without any descriptive information.

Virus: Any entity recognized as a virus by the annotators such as “West Nile virus”, “Flanders virus” or “Western equine encephalitis”. This also includes any abbreviations indicated in text to represent this virus such as “rabies”, (“RV”). All virus families, genus and subspecies should also be tagged as a virus.

Instructions for tagging: Tag all components of the virus name as one tag. This includes the word virus, except for commonly identified viruses such as “rabies” and “influenza”.

Special Cases: Words describing the virus should not be included as part of the tag, with the exception of “avian influenza” and “swine influenza”.

GeneName: Any entity recognized as a gene by the annotators. Some examples include “E1 envelope glycoprotein” gene, “N” gene, and “Matrix” gene.

Instructions for tagging: Include all components of the gene in a single tag, excluding the word gene. If an entity is tagged as a GeneName entity earlier in the text, and the context does not clearly indicate that the entity is no longer referring to a gene, then it should be tagged as a gene as well.

Appendix D

Description of Data Sources and Software

A number of freely available databases and open source systems were used to complete this project. This section provides a concise overview of each of these resources.

ISO 3166-1 alpha-2 Table: ISO (International Organization for Standardization) is the principal designer of voluntary International Standards in the world. The standard developed by them for representing countries is called ISO 3166-1. This standard allows representation of country names by a two-letter code (alpha-2), a three-letter code (alpha-3) and a three-digit numeric code (numeric-3) respectively. The table of alpha-2 country codes is freely available for non-commercial purposes and currently contains codes for 249 distinct countries.

GeoNames Database: The GeoNames database is a freely available, manually curated database that contains geospatial data for over 10 million locations on earth. Some of its most important data sources include National Geospatial-Intelligence Agency's (NGA) and the U.S. Board on Geographic Names, U.S. Geological Survey Geographic Names Information System and Ordnance Survey OpenData. Users also have the option to manually edit and add places using a wiki interface. Each location in this database is assigned a specific feature code corresponding to one of nine distinct feature classes. The database also contains the ISO country code of the country where each place is located.

BANNER: BANNER is an open-source named-entity recognition system, designed principally for biomedical text. It is implemented using conditional random fields and incorporates a rich feature set consisting primarily of orthographic, morphological and shallow-syntax features. It also provides users with the option of including a dictionary. BANNER was evaluated using 5x2 cross validation on the training corpus for BioCreative 2 Gene Mention Task and found to perform better than two of the existing, freely-available, state-of-the-art systems, ABNER and LingPipe, with a precision, recall and f-measure of 82.39%, 76.21% and 79.18% respectively. It is currently one of the most widely cited NER systems in biomedical literature for gene tagging**.**

LINNEAUS: LINNEAUS is a well-known, open-source species name identification and normalization system which uses a dictionary-based approach. The species dictionary used by this system was constructed using the NCBI taxonomy which contains 386,108 species names along with 116,557 genera and other higher-order taxonomic units. Different heuristics were used for successful disambiguation of overlapping mentions. The system was evaluated on various corpora by comparing both document level tags (such as MESH tags) and mention level tags (tags indicating the exact location of each mention) with those produced by LINNEAUS. On a manually annotated corpus of 100 full-text PubMed articles, LINNEAUS had a recall and precision of 94% and 97% respectively at mention level.

Stanford SUTime: SUTime is a temporal tagger created by the Stanford Natural Language Processing Group. It uses regular expressions to recognize and normalize time expressions and includes TIMEX3 tags in its annotations. TIMEX3 is a part of TimeML, a widely used formal specification language for events and temporal expressions. SUTime was evaluated on the TemEval-2 Task and found to be the best performer in the recognition of temporal expression extents with a token level precision, recall and f-measure of 0.88, 0.96 and 0.92 respectively.

Creation and Validation of an EMR-based Algorithm for Identifying Major Adverse Cardiac Events while on Statins

Wei-Qi Wei[1*], MMed PhD; Qiping Feng[2*], PhD; Peter Weeke, MD[2]; William Bush, PhD, MS[3]; Magarya S. Waitara[2], BA; Otito F. Iwuchukwu [2], PhD; Dan M. Roden[2,5,6], MD; Russell A. Wilke[4], MD PhD; Charles M Stein, M.B.Ch.B. [2]; and Joshua C. Denny[1], MD MS

1 Department of Biomedical Informatics, Vanderbilt University, Nashville, TN
2 Division of Clinical Pharmacology, Vanderbilt University School of Medicine, Nashville, TN
3 Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN
4 Department of Internal Medicine, Sanford Healthcare, Fargo, ND
5 Oates Institute for Experimental Therapeutics, Vanderbilt University, Nashville, TN
6 Office of Personalized Medicine, Vanderbilt University, Nashville, TN
*contributed equally

Abstract:

Statin medications are often prescribed to ameliorate a patient's risk of cardiovascular events due in part to cholesterol reduction. We developed and evaluated an algorithm that can accurately identify subjects with major adverse cardiac events (MACE) while on statins using electronic medical record (EMR) data. The algorithm also identifies subjects experiencing their first MACE while on statins for primary prevention. The algorithm achieved 90% to 97% PPVs in identification of MACE cases as compared against physician review. By applying the algorithm to EMR data in BioVU, cases and controls were identified and used subsequently to replicate known associations with eight genetic variants. We replicated 6/8 previously reported genetic associations with cardiovascular diseases or lipid metabolism disorders. Our results demonstrated that the algorithm can be used to accurately identify subjects with MACE and MACE while on statins. Consequently, future e studies can be conducted to investigate and validate the relationship between statins and MACE using real-world clinical data.

Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide. Recent mortality data show that CVD accounted for 32.8% of all deaths in the U.S.[1] Many randomized clinical trials (RCTs) have shown that HMG-CoA reductase inhibitors ("statins") significantly reduce the frequency of major adverse cardiac events (MACE) in patients at risk.[2-7] Statins are one of the most commonly prescribed medications, and are generally well-tolerated.[8] Given their clinical importance, they have been a frequent focus of investigation in electronic medical records (EMRs). We sought to develop a highly accurate algorithm to enable study of statin efficacy, measured as MACE while on statins, in EMRs. This algorithm can be used for later clinical and genomic studies.

Since 2000, EMRs have been widely implemented through the U.S.[9] The deployment of EMRs not only improves patient care but also generates huge clinical practice-based datasets ideal for evaluating previous findings from randomized controlled trials (RCTs).[10-13] Although useful for research, EMR data often requires carefully constructed algorithms to accurately identify phenotypes for clinical and genomic study[10,14-16]; this is especially true for pharmacogenomic studies in the EMR, since they require knowledge of the temporal relationship between exposures and outcomes. Once accurate algorithms are identified, studies can be conducted to investigate relevant relationships, e.g., between statins and MACE, using real-world clinical data.

Background

MACE can be defined as cardiac death, nonfatal acute myocardial infarction (AMI), or target lesion revascularization. Previously, several investigators have explored the possibility of identifying MACE subjects using EMR data. In 1996, Pladevall et al., reported that the accuracy of using the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code 410 to identify definite MI was 92%.[17] Similarly,

Petersen et al. in 1999 found that the positive predictive value (PPV) of AMI codes in the primary position was 96%. In addition, they also reported that the sensitivity and specificity of Current Procedural Terminology (CPT) coding were, respectively, 96% and 99% for coronary catheterization, 95% and 100% for coronary artery bypass graft surgery, and 90% and 99% for percutaneous transluminal coronary angioplasty.[18] In 2002, Austin et al., examined the use of a discharge diagnosis of AMI and the PPV was 88%.[19] In 2004, Kiyota et al., additionally required hospitalization lasting at least 3 days. Their results reflected a slightly improved PPV of 94%.[20] Two recent studies, by Varas-Lorenzo et al. in 2008[21] and by Preciosa et al. in 2013[22], reported that ICD9-CM codes had a PPV of 95% and 96%, respectively. Generally, these results suggest that ICD-9-CM codes have been widely used for MACE subject identification and yield PPVs in the mid to high 90% range.[23] However, all these studies were performed on primary/secondary discharge codes only (thus representing inpatient-generated codes, which typically result from professional coders). Such information is not available for many deidentified EMR datasets, i.e. it may not be clear if a code is for the principal or discharge diagnosis. Thus, the approach of simply using ICD9-CM codes may not generalize to a broad clinical research setting. Another important issue is identifying first MACE events. The recognition of such events empowers researchers to evaluate the effectiveness of a treatment for either primary or secondary prevention of MACE, therefore, has a foreseeable and significant impact on clinical practice.

Recent studies have begun using EMR data for pharmacological studies. Drug response phenotypes can be challenging to identify accurately, as they require presence of a medication during the timing of an event. [24] In a recent paper, we described our methods for extracting information and constructing full dose-response curves for simvastatin and atorvastatin using EMR data. [25] Advanced techniques, e.g. natural language processing (NLP) and ontology, were used to retrieve medication and laboratory data from structured and unstructured EMRs. Other examples of pharmacological studies include pharmacogenetic studies of clopidogrel and *CYP2C19* variants, in which manual review was ultimately required to achieve PPV[26], and the affect of common variants with warfarin stable-dose international normalized ratios (INRs) , which was able to be performed entirely using informatics techniques.[27] Other clinical studies have used NLP, sometimes with laboratory data, to replicate known drug adverse events and suggest some others, though formal assessments of the PPV of each drug-event pair were not provided.[28,29]

In this manuscript, we introduce an algorithm to identify subjects with MACE while on statins from EMRs. We report its performance compared to manual chart review and a genetic validation study. Compared to other efforts, our algorithm involves all diagnosis codes as well as laboratory data and simple NLP, instead of just primary discharge codes; it also assesses concurrent statin use, and includes a determination of first MACE.

Methods

MACE Algorithm development:

We used commonly captured EMR data, including ICD9-CM codes, CPT codes, and laboratory test results to develop an approach to identify MACE. We used all diagnosis codes rather than primary discharge codes alone so that our approach would be widely generalizable.

We categorized a MACE event as either AMI or revascularization. Qualifying cases of AMI while on statins were required to have ≥2 AMI relevant ICD9-CM Codes (410.* or 411.*) within a 5-day window and an abnormal laboratory test (Table 1). An abnormal laboratory test was defined as either troponin ≥0.10 ng/ml or both creatinine kinase (CK) MB fraction to CK ratio≥3.0 and CK-MB ≥10.0 ng/mL. In addition, a statin must have been prescribed prior to the AMI event ≥180 days (Figure 1). We chose slightly higher thresholds than usual to ensure the accuracy of the algorithm. The duration of 180 days was chosen empirically to represent a time course for which a patient would have significant statin exposure before their event and to make it easier to ascertain whether the patient had remained on the medicine. Statins were either simvastatin (Zocor), fluvastatin(Lescol, Canef, Vastin), atorvastatin (Lipitor), pravastatin(Pravachol, Selektine), lovastatin(Mevacor), cerivastatin (Baycol, Lipobay), or rosuvastatin(Crestor). Medications were identified using records from electronic prescribing tools and processing of free text notes using MedEx[30].

For qualified subjects with an AMI while on statins, we identified individuals with 1st AMI events, as those with no AMI codes (410 - 412) prior to the qualifying statin exposure-AMI event and with no other MACE history defined by applying NLP on previous notes. We used the KnowledgeMap Concept Indexer (KMCI)[31,32], a general-purpose NLP engine, to parse a patient's notes. Any non-negated keywords found, including *AMI*, *MI*, *acute myocardial*

infarction, *myocardial infarction*, *CABG*, *coronary artery bypass*, *cypher*, *taxus*, *BMS*, *DES*, and *stent,* was considered as an indication of positive MACE history and thereby excluded as a subject.

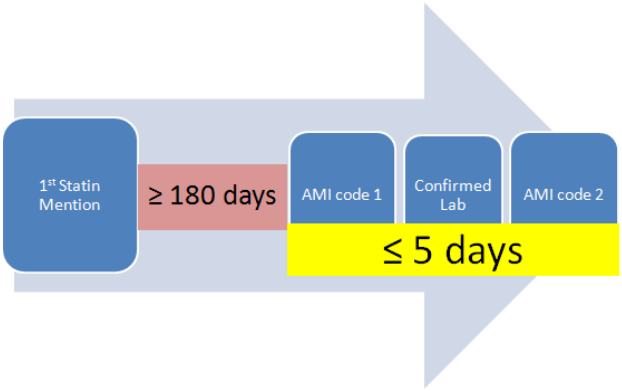


Figure 1. Overview of algorithm for determining AMI on statins

Revascularization includes percutaneous coronary intervention (PCI) and coronary artery bypass grafting (CABG). To be a qualified subject with revascularization while on statins, one must have a revascularization CPT code and a statin must be prescribed prior to the procedure ≥180 days (Table 1). The CPT codes that we used included coronary artery bypass (33533-33536, 33510-33523), angioplasty (92980-92982, 92984, 92995, 92996), and stent (C1874-C1877). Individuals with 1st revascularizations while on statins were those whom met the above criteria and had no revascularization CPT codes and no revascularization history found by NLP prior to the MACE on statin event.

Table 1. Algorithm for identifying subjects with MACE while on statins.

AMI on statin	• ≥ 2 AMI Codes (410.* or 411.*) within a 5-Day Window
	• Abnormal lab within the same time window defined by
	  ○ Troponin-I ≥ 0.10 ng/ml
	  ○ or Troponin-T ≥ 0.10 ng/ml,
	  ○ or CK-MB/CK ratio ≥ 3.0 and CK-MB ≥ 10.0 ng/mL
	• Statin prescribed prior to the AMI event ≥180 days

1st AMI on statin	• AMI on statin • No AMI codes (410 - 412) assigned before the AMI event • No MACE history defined by NLP
Revascularization while on statin	• Any CPT code for angioplasty, stent, or CABG • statin prescribed prior to the procedure ≥180 days
1st Revascularization while on statin	• Revascularization while on statin • No revascularization codes assigned before the AMI event • No MACE history defined by NLP

We similarly developed an algorithm to identify control subjects without MACE while on statin. We excluded patients with any AMI diagnosis or revascularization CPT codes, patients with previous history of AMI or revascularization defined by NLP. We also required controls to have had similar statin exposure in their EMRs matched with cases.

Manual chart review:
We applied the algorithm on BioVU individuals at Vanderbilt University Medical Center (VUMC) [33] to identify possible cases. In brief, BioVU links a de-identified image of the Vanderbilt EMR to DNA extracted from blood samples (obtained during routine clinical care and about to be discarded). Each record and associated DNA sample is linked by a unique identifier generated by a one-way hash function. The resource has been considered as containing data for nonhuman subjects in accordance with the provisions of Title 45 of the Code of Federal Regulations part 46, as have the individual research studies utilizing the resource.[33] As of 09/2013, BioVU contains > 170,000 unique individuals, including their dense longitudinal clinical records and associated blood samples.

From each category (AMI on statin, 1st AMI on statin, revascularization on statin, and 1st revascularization on statin), a group of 30 randomly selected cases was manually reviewed by two physicians. AMI on statin and 1st AMI on statin cases were reviewed by JCD, an internist. Revascularization on statin and 1st revascularization on statin cases were reviewed by PW, a cardiologist.

Genetic validation:
To further illustrate the application of our algorithm, we performed a genotype and phenotype association study, also by leveraging BioVU resources. The study population consisted of the first 7747 European–Americans accrued into BioVU. The only selection criteria were that they met the general conditions for eligibility for BioVU; no clinical inclusion or exclusion criteria were applied. These subjects have already been genotyped in previous studies.[34] In the current analysis, we identified 533 MACE cases and 2,642 MACE-free controls and compared the frequency of eight selected SNPs with previously known associations with cardiovascular diseases or lipid metabolism among cases and controls (**Table 3**): rs1045642 [pharmacogenetic predictors of lipid-lowering response to atorvastatin][35], rs440446 [ApoE gene, Variations in ApoE affect cholesterol metabolism, which in turn alter risk of heart disease and in particular a heart attack or a stroke][36], rs2200733[atrial fibrillation (AF) and ischemic stroke][37,38], rs405509 [CAD][39], rs1333049 [CAD][40,41], rs1800795 [CAD][42], rs1800888 [MACE after PCI][43], and rs1048101[hypertension] [44]. SNPs were genotyped in DNA samples from these subjects. Genotyping was conducted using commercial Taqman Allelic discrimination assays available through Applied Biosystems, Inc. (ABI, Foster City, CA, USA). The case-control analyses was performed using PLINK, a free, open-source genetic analysis toolset (http://pngu.mgh.harvard.edu/~purcell/plink/).[45] This platform was selected based on its efficiency, flexibility and ease of application. The primary outcome of this validation was to replicate these associations using our MACE algorithm.

Results

Table 2 summarizes manual chart review results that ranged from 90% to 97% positive predictive value (PPV) for MACE case identification. We observed some false positives that were caused by system coding errors, e.g. a "stent" code assigned for an esophageal stent placement. The algorithm performed well on identifying the 1st event (PPV ~90%). Some previous major events were missed because they happened long time ago (before 1990) and were not recorded in our current system.

Table 2. Results of manual chart review

Category	PPV
Any AMI event	96.67%
1st AMI event	96.67%
Any AMI while on Statin	90.00%
1st AMI event while on Statin	90.00%
Any revascularization event	96.55%
1st revascularization event	89.66%
Any revascularization while on Statin	96.55%
1st revascularization event while on Statin	89.66%

A total of 533 MACE cases and 2,642 MACE-free controls were identified from 7747 subjects of the demonstration cohort. Eight pre-selected SNPs were genotyped for all 3175 subjects. Variants with call rate less than 99% were removed from final analyses. Case-control analysis successfully replicated six out of the eight previously reported associations with cardiovascular diseases or lipid metabolism disorders.

The validation results were shown in **Table 3**. The strongest association was observed from the variant located in ABCB1 gene (*rs1045642*). This SNP— rs1045642, has already been proven to influence the body response to atorvastatin[35], therefore potentially affects our cardiovascular endpoint— MACE. Two SNPs (*rs440446*, *rs405509*) located in ApoE gene were replicated. Both of them play a critical role in cholesterol metabolism, which in turn affect the development of heart disease.[36,39] *Rs2200733* is another important cardiovascular relevant SNP that we replicated. Numerous studies have reported that it is strongly associated with CAD regardless of race.[37,38,46-50] We also validated the associations between MACE and two adrenergic receptor SNPs *rs1048101* and *rs1800888*. The former has been previously reported to be able to alter the alpha1-adrenergic receptor autoantibody production in

hypertensive patients [44] while the latter is associated with a more aggressive CAD and adversely affects prognosis in a study of 330 patients undergoing PCI[43].

Table 3. Association between eight SNPs previously reported to be associated with CV disease with MACE on statins in our population.

Chr.	SNP	Gene/Association	Minor Allele Frequency	*p-value*
7	rs1045642	ABCB1/predictors of lipid-lowering response to atorvastatin	0.472	***0.001***
8	rs1048101	ADRA1A/hypertension	0.457	***0.006***
19	rs440446	ApoE/cholesterol metabolism, heart disease, AMI, and stroke	0.348	***0.009***
4	rs2200733	4q25/atrial fibrillation(AF), ischemic stroke	0.119	***0.016***
19	rs405509	ApoE/CAD	0.474	***0.032***
5	rs1800888	ADRB2/MACE after undertaking PCI	0.012	***0.040***
9	rs1333049	CDKN2B/CAD	0.479	0.121
7	rs1800795	IL6/CAD	0.418	0.940

Discussion

In this paper, we reported a novel algorithm for use in EMRs to accurately identify cases with MACE and 1st MACE while on statin. The algorithm achieved 90% to 97% PPVs for the identification of MACE cases as compared to clinician review. By applying the algorithm to EMR data of demonstration cohort in BioVU, cases and controls were identified and used subsequently to replicate six out of eight associations with known genetic variants. Our results demonstrated that the algorithm can be used to accurately identify cases with MACE while on statins.

Acknowledgements

The authors would like to acknowledge funding by the American Heart Association (13POST16470018), NIH/NLM (R01 LM 010685-01A1), T32 GM007569, UL1RR0 24975 (Vanderbilt CTSA, please double check), PGPoP (5 U19 HL065962-11), and PARC (U19 HL069757).

Reference

1. Roger VL, Go AS, Lloyd-Jones DM, et al. Executive summary: heart disease and stroke statistics--2012 update: a report from the American Heart Association. *Circulation.* Jan 3 2012;125(1):188-197.
2. Baigent C, Keech A, Kearney PM, et al. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet.* Oct 8 2005;366(9493):1267-1278.
3. Delahoy PJ, Magliano DJ, Webb K, Grobler M, Liew D. The relationship between reduction in low-density lipoprotein cholesterol by statins and reduction in risk of cardiovascular outcomes: an updated meta-analysis. *Clinical therapeutics.* Feb 2009;31(2):236-244.
4. Hao PP, Chen YG, Wang JL, et al. Meta-analysis of the role of high-dose statins administered prior to percutaneous coronary intervention in reducing major adverse cardiac events in patients with coronary artery disease. *Clinical and experimental pharmacology & physiology.* Apr 2010;37(4):496-500.
5. Kim MC, Ahn Y, Cho KH, et al. Early statin therapy within 48 hours decreased one-year major adverse cardiac events in patients with acute myocardial infarction. *International heart journal.* 2011;52(1):1-6.
6. Ridker PM, Pradhan A, MacFadyen JG, Libby P, Glynn RJ. Cardiovascular benefits and diabetes risks of statin therapy in primary prevention: an analysis from the JUPITER trial. *Lancet.* Aug 11 2012;380(9841):565-571.
7. Sato H, Kinjo K, Ito H, et al. Effect of early use of low-dose pravastatin on major adverse cardiac events in patients with acute myocardial infarction: the OACIS-LIPID Study. *Circulation journal : official journal of the Japanese Circulation Society.* Jan 2008;72(1):17-22.
8. Maji D, Shaikh S, Solanki D, Gaurav K. Safety of statins. *Indian journal of endocrinology and metabolism.* Jul 2013;17(4):636-646.
9. Shea S, Hripcsak G. Accelerating the use of electronic health records in physician practices. *The New England journal of medicine.* Jan 21 2010;362(3):192-195.
10. Wilke RA, Xu H, Denny JC, et al. The emerging role of electronic medical records in pharmacogenomics. *Clin Pharmacol Ther.* Mar 2011;89(3):379-386.
11. McCarty CA, Wilke RA. Biobanking and pharmacogenomics. *Pharmacogenomics.* May 2010;11(5):637-641.
12. Ritchie MD, Denny JC, Crawford DC, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *American journal of human genetics.* Apr 9 2010;86(4):560-572.
13. Pulley JM, Denny JC, Peterson JF, et al. Operational implementation of prospective genotyping for personalized medicine: the design of the Vanderbilt PREDICT project. *Clin Pharmacol Ther.* Jul 2012;92(1):87-95.
14. Wei WQ, Leibson CL, Ransom JE, Kho AN, Chute CG. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *International journal of medical informatics.* Apr 2013;82(4):239-247.
15. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. *Science translational medicine.* Apr 20 2011;3(79):79re71.
16. Wei WQ, Leibson CL, Ransom JE, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *Journal of the American Medical Informatics Association : JAMIA.* Mar-Apr 2012;19(2):219-224.

17. Pladevall M, Goff DC, Nichaman MZ, et al. An assessment of the validity of ICD Code 410 to identify hospital admissions for myocardial infarction: The Corpus Christi Heart Project. *International journal of epidemiology.* Oct 1996;25(5):948-952.
18. Petersen LA, Wright S, Normand SL, Daley J. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *Journal of general internal medicine.* Sep 1999;14(9):555-558.
19. Austin PC, Daly PA, Tu JV. A multicenter study of the coding accuracy of hospital discharge administrative data for patients admitted to cardiac care units in Ontario. *American heart journal.* Aug 2002;144(2):290-296.
20. 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. *American heart journal.* Jul 2004;148(1):99-104.
21. Varas-Lorenzo C, Castellsague J, Stang MR, Tomas L, Aguado J, Perez-Gutthann S. Positive predictive value of ICD-9 codes 410 and 411 in the identification of cases of acute coronary syndromes in the Saskatchewan Hospital automated database. *Pharmacoepidemiology and drug safety.* Aug 2008;17(8):842-852.
22. Coloma PM, Valkhoff VE, Mazzaglia G, et al. Identification of acute myocardial infarction from electronic healthcare records using different disease coding systems: a validation study in three European countries. *BMJ open.* 2013;3(6).
23. Cutrona SL, Toh S, Iyer A, et al. Design for validation of acute myocardial infarction cases in Mini-Sentinel. *Pharmacoepidemiology and drug safety.* Jan 2012;21 Suppl 1:274-281.
24. Roden DM, Xu H, Denny JC, Wilke RA. Electronic medical records as a tool in clinical pharmacology: opportunities and challenges. *Clin Pharmacol Ther.* Jun 2012;91(6):1083-1086.
25. Wei WQ, Feng Q, Jiang L, et al. Characterization of Statin Dose-response within Electronic Medical Records *Clin Pharmacol Ther.* 2013.
26. Delaney JT, Ramirez AH, Bowton E, et al. Predicting clopidogrel response using DNA samples linked to an electronic health record. *Clin Pharmacol Ther.* Feb 2012;91(2):257-263.
27. 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.* Jul-Aug 2011;18(4):387-391.
28. Tatonetti NP, Denny JC, Murphy SN, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. *Clin Pharmacol Ther.* Jul 2011;90(1):133-142.
29. LePendu P, Iyer SV, Bauer-Mehren A, et al. Pharmacovigilance using clinical notes. *Clin Pharmacol Ther.* Jun 2013;93(6):547-555.
30. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association : JAMIA.* Jan-Feb 2010;17(1):19-24.
31. Denny JC, Peterson JF, Choma NN, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *Journal of the American Medical Informatics Association : JAMIA.* Jul-Aug 2010;17(4):383-388.
32. Denny JC, Bastarache L, Sastre EA, Spickard A, 3rd. Tracking medical students' clinical experiences using natural language processing. *Journal of biomedical informatics.* Oct 2009;42(5):781-789.
33. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther.* 2008;84(3):362-369.
34. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics.* May 1 2010;26(9):1205-1210.

35. Rosales A, Alvear M, Cuevas A, Saavedra N, Zambrano T, Salazar LA. Identification of pharmacogenetic predictors of lipid-lowering response to atorvastatin in Chilean subjects with hypercholesterolemia. *Clinica chimica acta; international journal of clinical chemistry.* Feb 18 2012;413(3-4):495-501.

36. Andreotti G, Menashe I, Chen J, et al. Genetic determinants of serum lipid levels in Chinese subjects: a population-based study in Shanghai, China. *European journal of epidemiology.* 2009;24(12):763-774.

37. Gretarsdottir S, Thorleifsson G, Manolescu A, et al. Risk variants for atrial fibrillation on chromosome 4q25 associate with ischemic stroke. *Annals of neurology.* Oct 2008;64(4):402-409.

38. Virani SS, Brautbar A, Lee VV, et al. Usefulness of single nucleotide polymorphism in chromosome 4q25 to predict in-hospital and long-term development of atrial fibrillation and survival in patients undergoing coronary artery bypass grafting. *The American journal of cardiology.* May 15 2011;107(10):1504-1509.

39. Fredriksson J, Anevski D, Almgren P, et al. Variation in GYS1 interacts with exercise and gender to predict cardiovascular mortality. *PloS one.* 2007;2(3):e285.

40. Karvanen J, Silander K, Kee F, et al. The impact of newly identified loci on coronary heart disease, stroke and total mortality in the MORGAM prospective cohorts. *Genetic epidemiology.* Apr 2009;33(3):237-246.

41. Buysschaert I, Carruthers KF, Dunbar DR, et al. A variant at chromosome 9p21 is associated with recurrent myocardial infarction and cardiac death after acute coronary syndrome: the GRACE Genetics Study. *European heart journal.* May 2010;31(9):1132-1141.

42. Antonicelli R, Olivieri F, Bonafe M, et al. The interleukin-6 -174 G>C promoter polymorphism is associated with a higher risk of death after an acute coronary syndrome in male elderly patients. *International journal of cardiology.* Sep 1 2005;103(3):266-271.

43. Piscione F, Iaccarino G, Galasso G, et al. Effects of Ile164 polymorphism of beta2-adrenergic receptor gene on coronary artery disease. *Journal of the American College of Cardiology.* Oct 21 2008;52(17):1381-1388.

44. Sun YX, Liao YH, Zhu F, et al. [Association between ADRA1A gene polymorphism and autoantibodies against the alpha1-adrenergic receptor in hypertensive patients.]. *Zhonghua xin xue guan bing za zhi.* Oct 2008;36(10):883-887.

45. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics.* Sep 2007;81(3):559-575.

46. Mohanty S, Santangeli P, Bai R, et al. Variant rs2200733 on chromosome 4q25 confers increased risk of atrial fibrillation: evidence from a meta-analysis. *Journal of cardiovascular electrophysiology.* Feb 2013;24(2):155-161.

47. Henningsen KM, Olesen MS, Haunsoe S, Svendsen JH. Association of rs2200733 at 4q25 with early onset of lone atrial fibrillation in young patients. *Scandinavian cardiovascular journal : SCJ.* Dec 2011;45(6):324-326.

48. Wnuk M, Pera J, Jagiella J, et al. The rs2200733 variant on chromosome 4q25 is a risk factor for cardioembolic stroke related to atrial fibrillation in Polish patients. *Neurologia i neurochirurgia polska.* Mar-Apr 2011;45(2):148-152.

49. Goodloe AH, Herron KJ, Olson TM. Uncovering an intermediate phenotype associated with rs2200733 at 4q25 in lone atrial fibrillation. *The American journal of cardiology.* Jun 15 2011;107(12):1802-1805.

50. Lee KT, Yeh HY, Tung CP, et al. Association of RS2200733 but not RS10033464 on 4q25 with atrial fibrillation based on the recessive model in a Taiwanese population. *Cardiology.* 2010;116(3):151-156.

Efficiently mining Adverse Event Reporting System for multiple drug interactions

Yang Xiang, PhD[1], Aaron Albin, BS[1,2], Kaiyu Ren, BS[1,2], Pengyue Zhang, MS[4], Jonathan P. Etter, PhD[3], Simon Lin, MD[5], Lang Li, PhD[4]
Department of [1]Biomedical Informatics and [2]Computer Science and Engineering and [3]Division of Medicinal Chemistry & Pharmacognosy, The Ohio State University, Columbus, OH 43210;
[4]Center for Computational Biology and Bioinformatics, Indiana University, Indianapolis, IN 46202
[5]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI 54449

Abstract

Efficiently mining multiple drug interactions and reactions from Adverse Event Reporting System (AERS) is a challenging problem which has not been sufficiently addressed by existing methods. To tackle this challenge, we propose a FCI-fliter approach which leverages the efforts of UMLS mapping, frequent closed itemset mining, and uninformative association identification and removal. By applying our method on AERS, we identified a large number of multiple drug interactions with reactions. By statistical analysis, we found most of the identified associations have very small p-values which suggest that they are statistically significant. Further analysis on the results shows that many multiple drug interactions and reactions are clinically interesting, and suggests that our method may be further improved with the combination of external knowledge.

Introduction

It is well understood that adverse drug reactions may pose serious health concerns on patients. The situation becomes more complicated when two or more drugs are taken together. Interactions between multiple drugs may yield additional reactions than taking them separately. To monitor the adverse drug reactions, the US Food and Drug Administration built an Adverse Event Reporting System (AERS), a post-marketing drug safety surveillance database which contains adverse reports from various sources.

However, AERS is essentially a large collection of drug reaction reports. A report involving multiple drugs and reactions does not necessarily indicate a causal relationship between them. In fact, records in AERS come from multiple sources coded as "Foreign", "Study", "Literature", "Consumer", "Health Professional", etc. It is not clear whether all sources produce similar accurate reports to AERS. Thus, mining such a large data for causative adverse drug reactions poses a major challenge in drug safety studies.

The existing work on AERS data mining and analysis mainly focuses on using statistic approaches. Some studies identify the reactions caused by one drug, or the drug-drug interactions between two drugs, using statistical approaches such as Bayesian methods [1] [2] and propensity score matching [3]. Some studies focus on the analysis of a few specific adverse reactions [4] or a few drug-drug interaction pairs [5]. In [2], the authors also extend the self-controlled case series (SCCS) to analyze multiple drug interactions. However, these methods did not answer the question of how to efficiently discover multiple drug interactions, i.e., drug-drug interactions that involve two or more drugs. There are many reports in AERS involving more than 2 drugs.

To tackle this challenge, Harpaz et al. [6] used association rules mining technique to find frequent patterns. A frequent pattern (a.k.a., frequent itemset) in AERS is a set of drugs and reactions that appear in at least k reports, where k is an adjustable parameter that is known as minimum support. The lower k is, the more patterns will be found and thus more computational time is needed. However, using frequent pattern mining has two major limitations.

First, it is computationally very costly. If a pattern is frequent, then all its sub patterns are frequent and should be outputted under the same support level k. A pattern with length x will have 2^x sub patterns (including the empty pattern and itself). This implies that it is computationally intractable to find a lengthy pattern because the number of sub patterns is exponential to its length. The counter measurement is to increase k or limit the output pattern size. But by doing this, we will miss a large volume of lengthy patterns and low support patterns. In [6], authors use

50, a quite high support level for mining AERS, and obtained only 2603 itemsets.

Second, the association rules suggested by frequent patterns are not sufficient to support the causative relationships between drug interactions and reactions. For example, if ($drug_A$, $drug_B$, $reaction_A$, $reaction_B$) is a frequent itemset, we cannot conclude that it is supportive evidence that the interaction of $drug_A$ and $drug_B$ leads to the $reaction_A$ and $reaction_B$. It may be caused by the facts that (1) $drug_A$ causes $reaction_A$; $drug_B$ causes $reaction_B$, $drug_A$ and $drug_B$ are often taken together.

Given the above challenging background, in this work we propose a very efficient mining method based on UMLS mapping, Frequent Closed Itemset Mining and filtering (FCI-filter) for mining multiple drug interactions from AERS. Our method efficiently finds a large number of multiple drug interactions and effectively prunes out uninformative patterns. It is important to point out that in this work we do not target on finding causative relationships between drug interactions and reactions, but on finding informative associations by eliminating associations that are not sufficient to support causative relationships.

Methods

UMLS Mapping

A drug or a reaction may have different names in the AERS, for example: Alpha Lipoic Acid is also known as ALA or Lipoic Acid. In many cases a drug name in AERS not only includes the drug but also its dosage. Therefore, it is not accurate to build a transactional database based on the drug or reaction names in AERS. To tackle this issue, we map each drug or reaction name to a UMLS concept, by LDPMap [7]. The UMLS is a very comprehensive collection of medical terms from various sources, such as HUGO, SNOMED CT, RxNorm, ICD9, MedDRA, etc. The RxNorm contains a large collection of drug names and has been successfully used in [6] for mapping drug names. The MedDRA was used for coding reactions in AERS. In the UMLS, a medical term may have various synonyms and may appear in more than one source, but it has only one unique identifier known as a CUI. In [7], we designed a layered dynamic programming mapping method (LDPMap) to effectively find a best matching UMLS CUI for any input of medical term. We have known that LDPMap is much more accurate in mapping medical terms to the UMLS than the UMLS Metathesaurus Browser [8] and MetaMap [9]. Here, we utilize LDPMap to map each drug and reaction to a UMLS CUI. In order to increase the accuracy, dosage related characters such as "oz", "ml" and "mg" in drug names were removed before applying LDPMap. After applying LDPMap on the AERS data of 2012q3, we obtained 10297 unique drugs and 6838 unique reactions, and built a transactional database AERS_tdb containing 134508 records.

Frequent Closed Itemset Mining

In data mining, a closed itemset is defined as an itemset which does not have a superset that has the same support as this itemset, and a frequent closed itemset is an itemset that is both closed and frequent. By using the concept of closed itemset, we will be able to eliminate the problem of enumerating exponential numbers of subsets. For example, if $drug_A$, $drug_B$, $reaction_A$, $reaction_B$ is a frequent closed itemset, then we do not need to output any of its subsets (such as $drug_A$, $reaction_A$) unless such a subset appears in a record that does not contain all items of $drug_A$, $drug_B$, $reaction_A$, $reaction_B$. Thus, we can see that by using the concept of frequent closed itemset, it is possible to significantly reduce the computational cost and eliminate the output of redundant information.

In this study, we use MAFIA [10], an efficient frequent closed itemset mining tool, to mine frequent closed itemset in AERS_tdb, with support level set to be 0.00005, which implies that any closed itemset appearing in 6.7254 or more records in AERS_tdb will be outputted. As a result, we obtained 4811379 frequent closed itemsets. Since we are interested in drug reaction relationships, we removed any itemset that contains only drugs or only reactions, and finally we got 1903630 itemsets containing both drugs and reactions. This is several orders of magnitude larger than the 2603 items obtained in [6]. In addition, we observed that the maximum number of drugs contained in one itemset is 20. This suggests that these 20 drugs are often taken together and with common reactions.

Uninformative Association Identification and Removal

As mentioned above, the association rules suggested by frequent closed itemsets are not equivalent to the causative relationships between drug interactions and reactions. An itemset is not sufficient to support a causative relationship if its items and supporting transactions (i.e., transactions containing these items) can be obtained from the interaction of other itemsets and their supporting transactions. In this case, this itemset is considered uninformative. Formally, Let I denote an itemset, and T denote the complete set of transactions containing this itemset. We have the following rule:

Rule 1: I is not sufficient to support causative relationships if there exist a list of itemset-transaction pairs $I_1 \times T_1$, $I_2 \times T_2$, … $I_n \times T_n$, $I = I_1 \cup I_2 \ldots \cup I_n$ and $T = T_1 \cap T_2 \ldots \cap T_n$ such that none of T_1 , $T_2 \ldots, T_n$ is equal to T.

In other words, if we view an itemset and its supporting transactions as a block, the above interaction can be described as a "block horizontal union" [11]. Thus, an itemset is not sufficient to support causative relationships if its block can be obtained by a block horizontal union on other blocks with different transaction sets. Here is an example:

$drug_A$, $reaction_A$, appears in and only in records 1, 3, 5

$drug_B$, $reaction_B$, appears in and only in records 1, 2, 5

$drug_A$, $drug_B$, $reaction_A$, $reaction_B$ appears in and only in records 1, 5.

Then $drug_A$, $drug_B$, $reaction_A$, $reaction_B$ is not sufficient to support a causative relationship such that the interaction of $drug_A$ and $drug_B$ causes $reaction_A$ and $reaction_B$, because this relationship is a logical result of taking both drugs together.

However, if in the above, $drug_A$, $reaction_A$ appears in and only in records 1, 5, then we cannot judge $drug_A$, $drug_B$, $reaction_A$, $reaction_B$ as "not sufficient to support a causative relationship".

In the following, we will use the above rule to eliminate frequent closed itemsets that are not sufficient to establish a causative relationship. Interestingly, we find that block interaction is not necessary for frequent closed itemsets and rule 1 can be simplified as:

Rule 2: A frequent closed itemset I is not sufficient to support causative relationships if there exist a list of frequent closed itemsets I_1, I_2, … I_n where $I = I_1 \cup I_2 \ldots \cup I_n$.

This is because for frequent closed itemsets, if $I = I_1 \cup I_2 \ldots \cup I_n$, we can conclude that for $T = T_1 \cap T_2 \ldots \cap T_n$, none of T_1 , $T_2 \ldots$, T_n is equal to T. Othewise, if one of the transaction set, say T_k, is equal to T, then it is a contradiction to the assumption that I_k is a closed itemset, because in this case $I_k \cup I$ would be a superset of I_k with the same support as I_k.

Next we will design an efficient filtering algorithm based on Rule 2. For an itemset I with p drugs, if $I = I_1 \cup I_2 \ldots \cup I_n$, we can observe that for any I_k $(1 \leq k \leq n)$, it must not contain more than p drugs. Thus, the filtering algorithm does not need to consider all itemsets in order to decide whether an itemset needs to be filtered out. We organize itemsets into groups by the number of drugs they contains. Let IS_k denote the itemset with k drugs, our filtering algorithm can be summarized by the following pseudo code:

```
Algorithm FCI-filter (IS1, IS2,…, ISm)
1:  for i=1:m
2:        for each itemset X in IS1∪…∪ISi
3:              for each itemset Y in ISi
                    if X⊂ Y
                      mark covered items in Y;
                    endif
                endfor
4:        endfor
          for each itemset Y in ISi
            if all items in Y are marked
                  remove Y;
            endif
          endfor
5:  endfor
9:  return IS1, IS2,…, ISm
```

By applying **FCI-Filter** to the 20 frequent closed itemsets mined from AERS_tdb, we filtered out 654484 frequent closed itemsets and kept 1249146 frequent closed itemsets as the candidate associate rules.

Statistical validation

We use the following statistical method to validate the filtered itemsets. Assume the counts for taking drug(s) and have reaction(s) follows a Poisson distribution. For any drug(s) and reaction(s), we will have the following frequency:

Total cases: N

Taking drug(s): a

Have reaction(s): b

If the drug(s) will not affect the rate of having reaction(s), the expected counts of taking drug(s) and having reaction(s) would be $\mu = b \times \frac{a}{N}$, as $\frac{a}{N}$ is the portion of people taking drug.

The P-value is based on the observed counts of taking drug(s) and having reaction(s) denoted by X and its expectation μ, which is $P(X > \mu)$, $X \sim Pois(\mu)$.

Results

By applying UMLS mapping and Frequent Closed Itemset Mining, we obtained a large number of

itemsets of drug interactions and reactions (Table 1). After applying algorithm FCI-Filter, we removed a significant amount of itemsets that are insufficient to support causative relationships (Table 1).

Number of drugs	Itemsets before filtering	Itemsets after filtering
1	1246948	48033
2	543037	1320
3	99755	144
4	11238	33
5	1231	14
6	267	12
7	155	9
8	100	3
9	83	3
10	57	2
11	42	1
12	43	1
13	57	0
14	96	0
15	139	1
16	159	0
17	135	1
18	70	2
19	17	0
20	1	0

Table 1. Summary of results of Frequent closed mining and frequent closed itemset filtering on AERS_tdb.

We subjected the itemsets (i.e., drug interactions and reactions) after filtering in Table 1 to statistical validation, and found that most itemsets have very significant low p-values (Figure 1). In addition, for drug counts greater than 10, p-value histogram (Figure 2) is similar to Figure 1, which further confirms the effectiveness of our drug interaction mining approach.

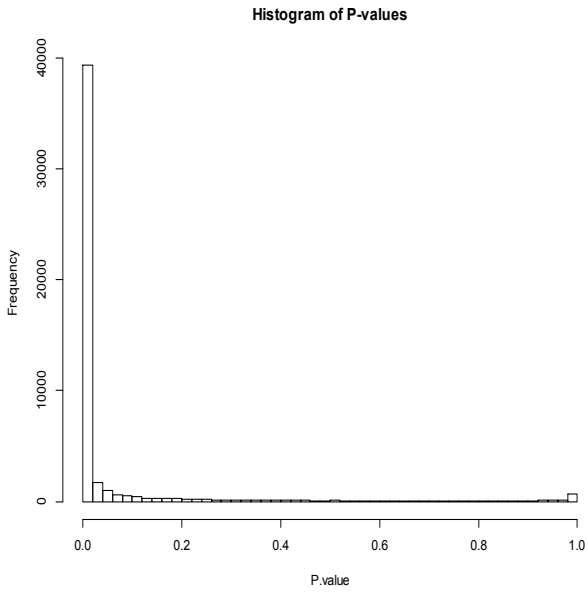


Figure 1. P-value histogram for all itemsets after filtering

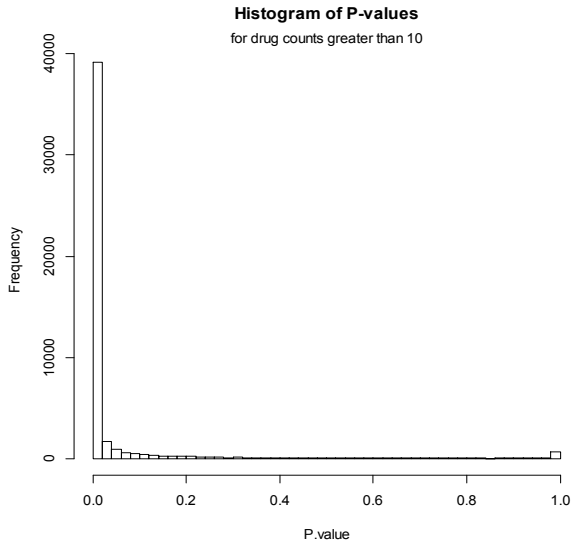


Figure 2. P-value histogram for drug counts greater than 10

Discussions

A clinical evaluation of the data mining results reveals some interesting findings as listed in Table 2.

Case	Drugs	Adverse Event
1	ARIPIPRAZOLE |CITALOPRAM HYDROBROMIDE |MIRTAZAPINE	CARDIAC FAILURE CONGESTIVE |CONGESTIVE CARDIOMYOPATHY
2	DULOXETINE HYDROCHLORIDE |MIRTAZAPINE |RISPERIDONE	LIVER FUNCTION TEST ABNORMAL
3	ASPIRIN |BISOPROLOL FUMARATE |GLYBURIDE |MIGLITOL|ONON |PLAVIX	HYPOGLYCAEMIA
4	AMARYL |SITAGLIPTIN PHOSPHATE	HYPOGLYCAEMIA
5	BROMOCRIPTINE MESYLATE |CLARITHROMYCIN |KETOCONAZOLE	HYPOTENSION

Table 2. Interesting drug drug interactions and reactions.

For instance, Aripiprazole, Citalopram hydrobromide and Mirtazapine, the three antidepressants sometimes used in combination therapies, were found to be in association with adverse cardiovascular events (Case 1 of Table 2). This result is highly interesting, since the potential cardiovascular side effects of antidepressants and antipsychotics have long been under debate [12] [13]. Recently in 2011, the US Food and Drug Administration (FDA) announced that "Citalopram causes dose-dependent QT interval prolongation. Citalopram should no longer be prescribed at doses greater than 40 mg per day." Further clinical study of Aripiprazole, Citalopram hydrobromide and Mirtazapine is required to explore their association with adverse cardiovascular events.

In addition to the above findings, we also observed interesting interactions involving a good number of drugs. For example, the following interaction contains 7 drugs and many reactions:

Drugs:
AMINOPYRIDINE|DANTRIUM|GILENYA|LEVO CARNIL|PIROXICAM|TROSPIUM CHLORIDE|VESICARE|

Reactions:
ALANINE AMINOTRANSFERASE INCREASED | ASPARTATE AMINOTRANSFERASE INCREASED | BLOOD CREATININE INCREASED |BLOOD GLUCOSE INCREASED|BLOOD LACTATE DEHYDROGENASE INCREASED|BLOOD UREA INCREASED|BLOOD URIC ACID DECREASED| |HAEMOGLOBIN DECREASED |…(18 other reactions)

The actions of this combination of drugs along with the reported biochemical effects is interesting. Many of these drugs act on ion channels or receptors, and the diverse array of biochemical effects that they result in is overwhelming. They result in increased activities of alanine aminotransferase, aspartate aminotransferase and blood lactate dehydrogenase. They also result in increased concentrations of blood creatinine, glucose and urea, as well as decreased concentrations in hemoglobin and blood uric acid. Many of these outcomes can be partly accredited to abnormal kidney or liver function, but they along with the other associated symptoms make analyzing their overall effects quite complex. However, this type of data analysis can provide valuable pieces of information that can act as a starting point in order to investigate why this combination of drugs has the resulting effects.

Future work

We have demonstrated in the above that FCI-filter is very effective in identifying important multiple drug interactions and reactions. However, the clinical evaluation also suggests some future improvements of our data mining strategy. An integration of clinical knowledge outside of the AERS database can be helpful (Case 3, 4, and 5 of Table 2). For instance, in Case 5 of Table 2, the hypotension side effect of Bromocriptine (single drug) is not statistically revealed from the AERS data set, although it is well known clinically to cause potential hypotension. As such, external knowledge can make the filtering of the Frequent Closed Itemset Mining more effective.

Acknowledgement

The project described was partially supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Bibliography

[1] W. DuMouchel, "Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous," *The American Statistician,* vol. 53, no. 3, pp. 177-190, 1999.

[2] D. Madigan, P. Ryan, S. Simpson and I. Zorych, "Bayesian Methods in Pharmacovigilance," *BAYESIAN STATISTICS,* vol. 9, pp. 421-438, 2010.

[3] N. P. Tatonetti, "Data-Driven Prediction of Drug Effects and Interactions," *Science Translational Medicine,* vol. 4, p. 125ra31, 2012.

[4] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase and C. Friedman, "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions," *J Am Med Inform Assoc,* vol. 20, p. 413–419, 2013.

[5] J. S. Almenoff, W. DuMouchel, L. A. Kindman, X. Yang and D. Fram, "Disproportionality analysis using empirical Bayes data mining: a tool for the evaluation of drug interactions in the post-marketing setting," *pharmacoepidemiology and drug safety,* vol. 12, p. 517–521, 2003.

[6] R. Harpaz, H. S. Chase and C. Friedman, "Mining multi-item drug adverse effect associations in spontaneous reporting systems," *BMC Bioinformatics,* vol. 11, no. Suppl 9, p. S7, 2010.

[7] K. Ren, A. Lai, A. Mukhopadhyay, R. Machiraju, K. Huang and Y. Xiang, "Effectively processing medical term queries on the UMLS Metathesaurus by layered dynamic programming," to appear in *BMC Medical Genomics,* vol. 7 (TBC 2013 Supplementary), 2014.

[8] "UMLS Metathesaurus Browser," [Online]. Available: https://uts.nlm.nih.gov.

[9] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program," in *Proceedings of the AMIA Symposium*, 2001.

[10] B. Douglas, M. Calimlim and J. Gehrke, "MAFIA: A maximal frequent itemset algorithm for transactional databases," in *17th International Conference on Data Engineering*, 2001.

[11] R. Jin, Y. Xiang, H. Hong and K. Huang, "Block interaction: a generative summarization scheme for frequent patterns," in *Proceedings of the ACM SIGKDD Workshop on Useful Patterns*, 2010.

[12] T. Acharya, S. Acharya, S. Tringali and J. Huang, "Association of Antidepressant and Atypical Antipsychotic Use with Cardiovascular Events and Mortality in a Veteran Population," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy,* 2013.

[13] P. J. Goodnick, F. Parra and J. Jerry, "Psychotropic drugs and the ECG: focus on the QTc interval," *Expert opinion on pharmacotherapy,* vol. 3, no. 5, pp. 479--498, 2002.

Adapting a Natural Language Processing Tool to Facilitate Clinical Trial Curation for Personalized Cancer Therapy

Jia Zeng, PhD[1], Yonghui Wu, PhD[2], Ann Bailey, PhD[1], Amber Johnson, PhD[1], Vijaykumar Holla, PhD[1], Elmer V. Bernstam, MD, MSE[2], Hua Xu, PhD[2], Funda Meric-Bernstam, MD[1]
[1]Institute for Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX; [2]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX

Abstract

The design of personalized cancer therapy based upon patients' molecular profile requires an enormous amount of effort to review, analyze and integrate molecular, pharmacological, clinical and patient-specific information. The vast size, rapid expansion and non-standardized formats of the relevant information sources make it difficult for oncologists to gather pertinent information that can support routine personalized treatment. In this paper, we introduce informatics tools that assist the retrieval and curation of cancer-related clinical trials involving targeted therapies. Particularly, we adapted and extended an existing natural language processing tool, and explored its applicability in facilitating our annotation efforts. The system was evaluated using a gold standard of 539 curated clinical trials, demonstrating promising performance and good generalizability (81% accuracy in predicting genotype-selected trials and an average recall of 0.85 in predicting specific selection criteria).

Introduction

It is now affordable to sequence an individual patient's genome and design personalized cancer treatment plans that directly target the underlying molecular aberrations. Personalizing therapy requires identifying the molecular alterations that "drive" cancer development in an individual patient, as well as associations between specific genomic alterations and specific targeted therapies. This information is used to optimally match patients to approved drugs and ongoing clinical trials of investigational targeted therapies. The process involves review and analysis of biomedical literature and other resources that provide information about molecular biology, targeted therapies and clinical trials. This laborious, manual process does not scale.

At MD Anderson Cancer Center (MD Anderson), the Institute for Personalized Cancer Therapy (IPCT) is dedicated to providing personalized cancer therapy to our patients. As part of our daily operation, IPCT's decision support team is manually curating biomedical literature, databases of targeted therapies and clinical trials to assist our physicians with personalized therapy selection. To expedite this curation effort, our institute is developing an informatics infrastructure that applies automated (or semi-automated) tools to achieve the following goals: 1) to retrieve and analyze molecular, pharmacological, clinical and patient-specific information from biomedical literature, targeted therapy and clinical trial databases as well as electronic health records; 2) to represent them in a standardized format and integrate them into a knowledge repository that is easy for the curators to navigate; and 3) to offer interfaces that enable the physicians to easily retrieve and visualize high quality curated information. In this paper, we report our progress in constructing a curated knowledge base of cancer-related clinical trials that involve targeted therapies, specifically regarding the identification of genotype-selected clinical trials and the genes used as their selection criteria.

To provide informatics support to facilitate this, we must properly identify gene entities inside the trial documents. However, gene name ambiguity is prevalent and causes a serious problem for the computational programs that try to extract genomic information from text [1]. Several methods have been proposed to disambiguate gene mentions in the biomedical literature [2-5]. However, very little work has been done to address this issue for clinical trial documents. Wu et al. developed a system that used natural language processing (NLP) techniques to disambiguate the status of a genetic lesion mentioned in a clinical trial document [6]. Specifically Wu's system captured four features of a gene mention: 1) the contextual words and the associated information; 2) words that have dependency relationships to the gene symbol based on the dependency parse tree from Stanford Parser [7]; 3) words expressing negation status; and 4) section headers including "title", "summary" and "eligibility criteria". Using a training set of 4332 manually annotated sentences, Wu et al. constructed a support vector machine-based classifier to identify the status of a gene mention as belonging to one of the following nine categories: 1) Drug Class (e.g., **AKT** inhibitor MK-2206); 2) Gene Status Altered or Gene Status Not Altered (e.g., any **HER2** status); 3) Gene Status Altered (e.g., with

documented **BRAF** mutation); 4) Gene Status Not Altered (e.g., wild type **MET** status); 5) Gene Status Unknown (e.g., **KRAS** mutation unknown); 6) Alteration Detected or Not Detected (e.g., selecting patients' whose **HER2** status is measured); 7) Gene Status (e.g., to compare **MET** status); 8) Gene (e.g., **PIK3CA** is a gene involved in the PI3K/AKT pathway); and 9) English Word (e.g., these requirements have to be **met**). The system achieved a highest accuracy of 89.8%, demonstrating its applicability in the real-world task of clinical trial annotation.

In this study, we adapted Wu's system [6] and assessed its performance using trials curated by the MD Anderson Cancer Center IPCT decision support team as the gold standard. Our evaluation demonstrated the merit of applying the system to facilitate manual curation and also validated the generalizability of the existing NLP tool.

Methods

Our system consists of a clinical trial retrieval and preprocessing component, a gene recognition component and a gene mention disambiguation component adapted from Wu's system.

Clinical Trial Retrieval and Pre-processing – To assist IPCT's daily operation, we have developed a program that automatically retrieved and pre-processed potential targeted therapy clinical trials from Clinicaltrials.gov [8] and the MD Anderson clinical trial database [9]. Given a set of applicable targeted therapies, the program automatically expanded the drug names by including known aliases (based on NCI's drug dictionary [10]) and retrieved matching clinical trials from Clinicaltrials.gov via its RESTful API (by constructing the search term for the "Interventions" field using the list of drug names concatenated with Boolean operator OR and formulating a query URL accordingly). The criteria for a match include: 1) the trial has to be an ongoing study (recruiting, not yet recruiting or available for expanded access); 2) it has to mention the drug name/alias in either the intervention or title sections; and 3) it needs to be applicable to at least one cancer type. The program then parsed the trial records (in XML format) returned by Clinicaltrials.gov, extracted and pre-processed pertinent information, and stored them in a tabular format. The fields included in the reformatted records were: Unique Trial Identifier (NCTID), Drugs, Applicable Conditions, Broadly Categorized Conditions, Detailed Recruitment Status, Phase, Title, Inclusion Criteria, Exclusion Criteria, General Criteria, Sponsors/Collaborators, Locations, and Hyperlinked URL to the Trial Document. If a trial was conducted at MD Anderson, then additional MD Anderson-specific information including the PI's name, clinic, and MD Anderson recruitment status would be provided.

It is worth noting that the criteria for selecting or excluding patients were typically provided under the eligibility section of the trial document and our program further divided the text into subsections based upon word boundaries (Inclusion or Exclusion) so the fields of Inclusion Criteria and Exclusion Criteria could be auto-populated. When no such boundary was found, all the content under eligibility would be used to populate a field called General Criteria.

Gene Recognition Component – In the existing version of Wu's system, the primary focus was to automatically disambiguate a gene mention. To identify the gene entities, Wu and colleagues first applied a string matching technique to identify potential occurrences of gene mentions and then used domain experts' feedback (e.g., confirmation, rejection or suggestion of change) to finalize the component of gene recognition.

To evaluate the feasibility of recognizing gene mentions without human intervention, in our adapted system, we constructed a gene recognizer by modifying a component from IPCT's existing information retrieval (IR) pipeline. The IR pipeline used Lucene (a text search engine library) [11] to index any textual document repository. At query time, it could take a human gene symbol as input, automatically expand it to include the official gene name and known aliases as indicated by NCBI's Entrez gene database [12], and retrieve the matching documents. The IR pipeline could also highlight the matched terms in text. To recognize genes for the purpose of identifying genotype-selected trials and their selection criteria, we modified the IR pipeline by first indexing the inclusion, exclusion and general criteria of all the trials in the gold standard, then using a set of predefined genes as the query and labeling the matched gene mentions in the trial documents by their corresponding gene symbols. To maintain consistency with IPCT's priorities, for the predefined gene list, we used a set of 543 genes whose molecular abnormality can be detected by at least one of the four sequencing panels offered at MD Anderson: CMS46 (46 gene Ampliseq platform, Ion Torrent, LifeTechnologies, Carlsbad CA), T200 and T300 (MD Anderson in-house targeted exome sequencing research platforms), and CMS400 (409 gene Ion Proton platform, Life Technologies, Carlsbad CA).

Adaptation of the Gene Mention Disambiguator – The disambiguator reported by Wu and colleagues [6] was trained and tested using a 9-class categorization system. To make the tool applicable to IPCT's curation tasks, we made the following adaptation: if the 9-class disambiguator labeled a gene mention to be class 2, 3 or 6 and the occurrence of the gene mention was not inside the trial's exclusion criteria, then predicted the trial as genotype-selected and labeled the official symbol of the gene mention as a selection criterion.

Results

Manual Curation and Generation of the Gold Standard – MD Anderson Cancer Center IPCT decision support team routinely performs manual review of clinical trials to answer the following questions: 1) whether the trial is genotype-selected, i.e., selecting for patients who have specific molecular abnormalities (e.g., PIK3CA mutation, MET amplification); 2) for a genotype-selected trial, which genes are the selection criteria; 3) whether the trial is genotype-relevant, i.e., does the trial involve a targeted therapy that is applicable to treating patients with matching molecular profiles (e.g. targeting downstream signaling activated by a molecular alteration); and 4) for a genotype-relevant trial, alterations in which genes may be relevant. For trials that have multiple cohorts and a specific molecular alteration (e.g., HER2 amplification) only applies to one cohort, we annotated them as genotype-selected.

To facilitate this study, i.e., to identify genotype-selected trials and their gene selection criteria, we constructed a gold standard of 571 clinical trials manually annotated by the IPCT team, where 153 trials were genotype-selected and the rest (418 trials) were non-genotype-selected. Notably there was no overlap between these trials and those used to train the 9-class disambiguator. Using our gold standard as a testing set, we assessed the performance of the gene recognition component and the adapted disambiguation component respectively.

Gene Recognition Component – Of the 153 genotype-selected trials in our gold standard, our gene recognizer was able to correctly identify all genes annotated as selection criteria in 121 trials. In the remaining 32 trials, at least one gene was not properly recognized.

Adapted Gene Mention Disambiguation Component – To assess the performance of our adapted disambiguator, we excluded the 32 trials that were not properly labeled by the gene recognition component and constructed a test set from the gold standard which included 121 genotype-selected trials and 418 non-genotype-selected trials. The binary classifier predicted 193 genotype-selected trials and 346 non-genotype-selected trials, yielding an accuracy of 81%. Precision and recall were 0.55 and 0.88 respectively, yielding an F score of 0.68. With the understanding that recall was not perfect (15 trials were erroneously labeled as non-genotype-selected), we entertained the following hypothetical analysis: if our curators did not have to curate the trials that were predicted to be non-genotype-selected, they would have saved 346 minutes worth of man power (approximately 1 minute used for concluding a trial that is non-genotype-selected), which made up 83% of all the time spent on annotating non-genotype-selected trials in the gold standard.

We also evaluated the performance of the disambiguator in identifying genes that serve as selection criteria. The averages of precision, recall and F score were 0.69, 0.85 and 0.74 respectively. Overall, recall was higher than precision, which is consistent with our expectation, since for our task, a false negative (missed trial) is much worse than a false positive. Table 1 shows the performance of our system on nine genes that were annotated as the selection criteria for at least 5 clinical trials in the gold standard.

Discussion

In this paper, we presented an informatics system that facilitates the retrieval, analysis and curation of genotype-selected clinical trials to guide personalized cancer treatment. We have adapted and extended an existing NLP tool trained at a different institution and investigated its applicability to our curation tasks. Using IPCT's in-house curated clinical trials as the gold standard, we evaluated the performance of the modified system and observed promising results with an average accuracy of 81% in predicting genotype-selected trials and an average recall of 0.85 in predicting the genes that serve as selection criteria. To understand the limitations of our current system and to shed light in our future direction, we performed an error analysis of the components of gene recognition and disambiguation. In the ensuing text, we elaborate on our analysis and identify opportunities for improvement which will be explored in our future studies.

Gene Recognition Component – Our error analysis has revealed the following five reasons why the gene recognizer failed to identify all the gene mentions in 32 trials.

1) Translocation/fusion genes (e.g., EML4-ALK): the individual components of a fusion gene were properly identified yet their co-occurrence in this context was not tagged as a fusion gene. There were 7 trials that were incompletely labeled due to this. An enhancement that recognizes the pattern of translocation/fusion genes and tags them appropriately will overcome this limitation.

2) Gene mentioned outside of the eligibility criteria: in 8 trials, the selected genes were not mentioned in the eligibility criteria section, instead they were mentioned only in the title, summary or outcome description. While we still think it is reasonable to expect the trial document to be structured so critical information such as genes used

as selection criteria would occur in the eligibility criteria section, we can easily expand the sections to be analyzed to include title, summary and outcome.

3) Incomplete dictionary: to enable automatic query expansion given a gene symbol, we used NCBI's Entrez gene database as a dictionary to look up the genes' common aliases and official name. However from the 6 mislabeled trials we learned that such a dictionary would require some expansion. For instance, RAS is often used to refer to a family of genes encoding proteins in the RAS family, including NRAS, KRAS and HRAS, yet it is not included as an alias for any of these three genes. Similar observations have been made between the following alias/symbol pairs: RAF for **BRAF**, MEK for **MAP2K1** or **MAP2K2**, PDGFR for **PDGFRA** or **PDGFRB,** and CD79 for **CD79A** or **CD79B**. To overcome this problem, we will supplement the current dictionary by integrating additional resources that contain information about gene names (such as GeneCard [13]), and/or design rules that extrapolate an alias based upon the gene symbol. For instance, we can assume that a gene ending with a letter (e.g., CD79A) encodes a subunit of a protein (e.g., CD79) and automatically assign the root portion of the symbol (e,g., CD79) as an alias. In our current gene recognizer, we have already applied a similar rule for gene symbols ending with a digit (e.g., FGFR1/FGFR2/FGFR3/FGFR4) that belong to the same family (e.g., FGFR) and automatically included the family name as an alias.

4) Tokenizer: a commonly adopted convention for describing a point mutation is to place a delimiter (white space or dash) between a gene symbol and the point mutation (e.g., BRAF V600E or BRAF-V600E). However, our error analysis of 3 mislabeled trials revealed that in some trial documents, the delimiter was omitted (e.g., BRAFV600E). While this may not cause a problem for a gene recognizer that applies substring matching, it does introduce an issue to more sophisticated (and probably more efficient) information retrieval strategies that use a tokenizer which relies on such delimiters to identify word boundaries. To resolve this, we will customize the default tokenizer to recognize a pattern of point mutations (e.g., V600E) as a separate word.

5) Inferred by curators: there were 8 trials where the genes that were annotated as the selection criteria were not mentioned explicitly in the trial documents but were inferred by the curators based upon their domain knowledge. For example, the following text occurred in one trial that was mislabeled: "*with tumor mutations/amplifications in one of 3 genetic pathways (DNA repair, PI3K or RAS/RAF)*". While our tool was capable of recognizing the genes whose symbol/aliases are consistent with the pathway name, it was unable to infer what genes are associated with the DNA repair pathway. To achieve a perfect recall in this category, we would have to integrate pathway information into the gene recognition component.

We understand that the task of gene recognition and normalization is very sophisticated in its own right and the aforementioned analysis was not intended to be a comprehensive assessment of this component. Due to the scope of this study, we did not construct our gold standard in a manner that supports an in-depth evaluation of the gene recognizer. For future work, we plan to address this issue as well as exploring existing tools such as those evaluated at the BioCreative gene normalization competitions [14] (e.g. GenNorm [15], GeneTUKit [16] and IASL-IISR [17]).

Adapted Gene Mention Disambiguator Component – An examination of several trials that were erroneously classified by our disambiguation component revealed the following three common causes.

1) Negation boundary: to identify negation status, Wu et al. used a training set of a localized negation/assertion lexicon constructed by the domain experts based on their review of the clinical trial training set and applied a support vector machine based method to identify the negation cases and assertion cases in the stage-2 classification. The SVMs considered a rich set of features regarding negation status including: the direction of the negation words in relation to the gene symbols, the distance between the negation cues and the target gene symbol, and the punctuations. An error analysis is as follows. In situations like the following: "*Patients with histologically/cytologically confirmed advanced solid tumors with **FGFR1** or **FGFR2** amplification or **FGFR3** mutation, for which **no** further effective standard anticancer treatment exists*", the disambiguator successfully predicted the first and second gene mentions as "Gene Status Altered" but wrongly classified the third (FGFR3) to be "Gene Status Not Altered" because it is very close to a negation cue ("no"). In the future, we may explore the application of some existing negation analyzers to overcome this issue (e.g. NegEx by Chapman et al.[18] and Bejan et al's assertion analyzer [19]).

2) Drug class identification: in some cases, the disambiguation program failed to properly recognize that a gene is mentioned in the context of a drug. For example, the following cases were mislabeled as "Gene Status Altered": "*anti-EGFR antibody (cetuximab or panitumumab)*"; "*epidermal growth factor receptor (EGFR)*

inhibitor". Utilizing existing resources such as the UMLS that can help identify drug names could improve performance.

3) Disease acronym identification: some cancer types have acronyms that can be confused with an alias of a gene. For instance, papillary thyroid carcinoma is often abbreviated as PTC, which also happens to be an alias of the gene RET. In our current system, such ambiguity has not been taken into consideration. We plan to address this issue in our future work.

Conclusion

The existing NLP tool was generalizable. Informatics tools may partially automate the process of information gathering for the delivery of personalized cancer therapy. By conducting an error analysis, we identified several ways of further improving the performance of our system, which will be explored in our future studies.

Table 1. Performance evaluation of the adapted disambiguator from gene perspective.

Gene Symbol	# of Associated Genotype-Selected Trials in Gold Standard	Precision	Recall	F
BRAF	63	0.94	0.79	0.86
ERBB2	14	0.59	0.93	0.72
ALK	14	0.93	1.00	0.97
KRAS	11	0.65	1.00	0.79
PIK3CA	11	0.82	0.82	0.82
NRAS	8	0.88	0.88	0.88
PTEN	7	1.00	0.86	0.92
EGFR	6	0.45	0.83	0.59
MET	5	0.45	1.00	0.62

Acknowledgement

This study was supported in part by the MD Anderson Cancer Center Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy, NCI grant 1U01CA180964 and the Center for Clinical and Translational Science (NCATS UL1 TR000371). The authors would like to thank MyCancerGenome team at Vanderbilt University who provided annotation to the training set used by the original version of the gene mention disambiguator [6].

References

1. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 2005; 21(2): 248-56.
2. Schijvenaars BJ, Mons B, Weeber M, Schuemie MJ, van Mulligen EM, Wain HM, Kors JA. Thesaurus-based Disambiguation of gene symbols. BMC Bioinformatics. 2005; 6:149.
3. Xu H, Fan JW, Hripcsak G, Mendonca EA, Markatou M, Friedman C. Gene symbol disambiguation using Knowledge-based profiles. Bioinformatics. 2007; 23(8):1015-1022.

4. Farkas R. The strength of co-authorship in gene name disambiguation. BMC Bioinformatics. 2008; 9:69.
5. Stevenson M, Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. J Biomed Inform. 2010; 43(6):972-981.
6. Wu Y, Levy MA, Micheel CM, Yeh P, Tang B, Cantrell MJ, Cooreman SM, Xu H. Identifying the status of genetic lesions in cancer clinical trial documents using machine learning. BMC Genomics. 2012; 13:S21.
7. Klein D, Manning CD. Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. 2003:423-430.
8. Clinicaltrials.gov: an NIH funded registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. [http://clinicaltrials.gov/]
9. MD Anderson cancer center clinical trial registry. [http://www.mdanderson.org/patient-and-cancer-information/cancer-information/clinical-trials/clinical-trials-at-md-anderson/index.html]
10. NCI funded drug dictionary provides technical definitions and synonyms for drugs/agents used to treat patients with cancer or conditions related to cancer. [http://www.cancer.gov/drugdictionary]
11. McCandles M, Hatcher E, Gospodnetic O. Lucene in Action (2nd edition). Mannings Publications Co. 2010.
12. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2005; 33:D514-9.
13. Stelzer G, Harel A, Dalah A, Rosen N, Shmoish M, Iny Stein T, Sirota A, Madi A, Safran M and Lancet D. GeneCards: one stop site for human gene research. FISEB (ILANIT). 2008.
14. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, Hsu CN, Tsai RTH, Dai HJ, Okazaki N, Cho HC, Gerner M, Solt I, Agarwal S, Liu F, Vishnyakova D, Ruch P, Romacker M, Rinaldi F, Bhattacharya S, Srinivasan P, Liu H, Torii M, Matos S, Campos D, Verspoor K, Livingston KM, Wilbur WJ. The gene normalization task in BioCreative III. BMC Bioinformatics. 2011; 12(Suppl 8):S2.
15. GenNorm [http://ikmbio.csie.ncku.edu.tw/GN/]
16. Huang M, Liu J, Zhu X. GeneTUKit: a software for document-level gene normalization. Bioinformatics. 2011; 1(27):1032-3.
17. IASL-IISR Gene Mention/Normalization Tool. [http://sites.google.com/site/potinglai/downloads].
18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34:301-10.
19. Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion modeling and its role in clinical phenotype Identification. J Biomed Inform. 2013;46:68-74.

Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics

Ping Zhang, PhD, Fei Wang, PhD, Jianying Hu, PhD, Robert Sorrentino, MD
Healthcare Analytics Research Group, IBM T.J. Watson Research Center, New York, USA

Abstract

The rapid adoption of electronic health records (EHR) provides a comprehensive source for exploratory and predictive analytic to support clinical decision-making. In this paper, we investigate how to utilize EHR to tailor treatments to individual patients based on their likelihood to respond to a therapy. We construct a heterogeneous graph which includes two domains (patients and drugs) and encodes three relationships (patient similarity, drug similarity, and patient-drug prior associations). We describe a novel approach for performing a label propagation procedure to spread the label information representing the effectiveness of different drugs for different patients over this heterogeneous graph. The proposed method has been applied on a real-world EHR dataset to help identify personalized treatments for hypercholesterolemia. The experimental results demonstrate the effectiveness of the approach and suggest that the combination of appropriate patient similarity and drug similarity analytics could lead to actionable insights for personalized medicine. Particularly, by leveraging drug similarity in combination with patient similarity, our method could perform well even on new or rarely used drugs for which there are few records of known past performance.

Introduction

In contrast to the one-size-fits-all medicine, personalized medicine aims to tailor treatment to the individual characteristics of each patient. This requires the ability to classify patients into subgroups with predictable response to a specific treatment. The field of pharmacogenetics/pharmacogenomics has made important contributions to this problem for more than 50 years[1]. Ideally, personalized medicine will enable targeted prescription of any given treatment to only the likely responders, to avoid adverse reactions and expensive treatments in non-responders. Although there are already many examples of personalized medicine by leveraging genetics/genomics information in current practice[2], such information is not yet widely available in everyday clinical practice, and is insufficient since it only addresses one of many factors affecting response to medication.

With the tremendous growth of the adoption of EHR, various sources of clinical information (e.g., demographics, diagnostic history, medications, laboratory test results, vital signs) are becoming available about patients. Recently, some treatment comparison studies[3, 4] were conducted based on data from EHR of a cohort of clinically similar patients who received the treatments previously and whose outcomes were recorded. There are also some studies[5, 6] of combining clinical and genetics/genomics information in selecting optimal clinical treatments. Existing approaches using clinical information for personalized medicine rely on large amounts of real-world data regarding the target treatment itself, which may not be available for new drugs or rarely-used treatments.

Drug similarity analytics aims to find drugs which display similar pharmacological characteristics to the drug of interest. The similarity analytics is usually conducted based on one or more types of drug characteristics (e.g., chemical structures, biological targets, indications, side-effects, and gene expression profiles). Drug similarity analytics has been widely used in drug repositioning[7-9], drug side-effects prediction[10], drug-target interactions prediction[11], and drug-drug interactions prediction[12, 13] applications. This approach has been shown to deliver competitive or even better accuracy to more complex, feature-vector-based methods[9, 11] (e.g., support vector machines, random forests). In this study, we used drug similarity analytics to transmit EHR clinical information from well-studied drugs (i.e., drugs with many EHR records) to rarely-studied drugs (i.e., drugs with no or few EHR records).

Patient similarity analytics aims to find patients who display similar clinical characteristics to the patient of interest. The goal is to derive clinically meaningful distance metrics to measure the similarity between patients represented by their key clinical indicators. The resulting individualized insight of patient similarity analytics includes suggestions on how to manage care delivery to the patient (especially for patients has multiple diseases), and predictions of health issues that could arise in the future (because patients with similar characteristics had experienced such health issues). With the right patient similarity in place, patient similarity analytics have been used in the target patient retrieval[14], medical prognosis[15, 16], risk stratification[17, 18], and clinical pathway analysis[19] tasks.

In this study, we used patient similarity analytics to transmit EHR treatment information from training patients (i.e., patients with known effective treatments) to target patients (i.e., patients with no known effective treatment information).

In this paper, we construct a heterogeneous graph which includes two domains (i.e., patients and drugs) and encodes three relationships (i.e., patient similarity, drug similarity and patient-drug prior associations), and propose a heterogeneous label propagation algorithm which can be used to generate personalized drug recommendations by leveraging patient similarity and drug similarity analytics. To our best knowledge, the heterogeneous graph formulation of the EHR data has not been proposed in any previous literature. The label propagation model over heterogeneous graph by leveraging both patient similarity and drug similarity analytics is also significantly different from existing label propagation models.

Methodology

In this section we introduce the details of our method on how to combine patient and drug similarity analytics for personalized recommendations. There are three key components in our approach: drug similarity evaluation, patient similarity evaluation, and drug personalization.

Drug Similarity Evaluation. We used and compared chemical structure and drug target information to measure drug similarity. For chemical structure information, each drug was represented by an 881-dimensional binary profile whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Then we used the Tanimoto coefficient (TC), also known as the Jaccard index, to compute chemical structure similarities between all drug pairs. The TC between two vectors A and B is defined as the ratio between the number of features in the intersection to the union of both fingerprints: $TC(A,B) = |A \cap B|/|A \cup B|$. For drug target information, we collected all target proteins for each drug from DrugBank[20]. Then we calculated the pairwise drug target similarity between drugs d_x and d_y based on the average of sequence similarities of their target protein sets:

$$sim_{\text{target}}(d_x, d_y) = \frac{1}{|P(d_x)||P(d_y)|} \sum_{i=1}^{|P(d_x)|} \sum_{j=1}^{|P(d_y)|} SW(P_i(d_x), P_j(d_y))$$

where given a drug d, we presented its target protein set as $P(d)$; then $|P(d)|$ is the size of the target protein set of drug d. The sequence similarity function of two proteins SW was calculated as a Smith-Waterman sequence alignment score[21].

Patient Similarity Evaluation. We used co-occurring ICD9 diagnosis code information to measure patient similarity for simplicity and consistency purposes. In particular, we aggregated the longitudinal records of individual patients into a set of patient feature vectors, where each patient is a binary vector of ICD9 diagnosis categories. Then we used TC to compute similarities between all patient vectors.

Drug Personalization. As stated in the introduction, the basic question we want to answer for personalized medicine is "whether drug A is likely to be effective for specific patient B". To take into consideration the specific condition of patient B as well as the characteristics of drug A, we propose to leverage the information of the patients who are clinically similar to patient B as well as the drugs which are similar to drug A. Moreover, we also considered the prior associations between patients and drugs, which were measured by the TC between ICD9 diagnosis of patients and ICD9-format drug indications from MEDI database[22] (MEDI is an ensemble medication indication resource, which was created based on multiple commonly used medication resources by leveraging natural language processing techniques). In this way, we constructed a heterogeneous graph illustrated in Figure 1, which includes two domains (patients and drugs) and encodes three relationships (patient similarity, drug similarity and patient-drug prior associations). In the following we present a concrete heterogeneous label propagation algorithm to answer the question proposed at the beginning of this paragraph.

Suppose we have a set of patients $\boldsymbol{P}=\{p_1, p_2, ..., p_n\}$, where n is the number of patients with p_i representing the i-th patient, and a set of drugs $\boldsymbol{D}=\{d_1, d_2, ..., d_m\}$, where m is the number of drugs with d_j representing the j-th drug. Let $\boldsymbol{S}_p$ be the patient similarity matrix of size $n \times n$ with its (i,j)-th entry representing the similarity between p_i and p_j; $\boldsymbol{S}_d$ be the drug similarity matrix of size $m \times m$ with its (i,j)-th entry representing the similarity between d_i and d_j (in this study, the drug similarity comes from either chemical structure or drug target information source); and $\boldsymbol{R}$ be the patient-drug prior association matrix of size $n \times m$ with its (i,j)-th entry representing the association between p_i and d_j (in this study, the prior association comes from TC of patient diagnosis codes and drug indications). Then we can form a composite $(n+m) \times (n+m)$ patient-drug similarity matrix $\boldsymbol{A}$ by concatenating the three matrices as

$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{S}_p & \boldsymbol{R} \\ \boldsymbol{R}^T & \boldsymbol{S}_d \end{bmatrix}$. For each drug d, we constructed a corresponding effectiveness vector $\boldsymbol{y}=[y_1, y_2, \ldots, y_n, y_{n+1}, \ldots, y_{n+m}]^T$ where $y_k=1$ ($k=1,2,\ldots,n$) if d is an effective treatment for patient k, $y_k=1$ ($k=n+1,n+2,\ldots,n+m$) if d is the (k-n)-th drug, otherwise $y_k=0$. In this way, the effectiveness vector for each drug is just like a "label" vector on the heterogeneous graph shown on Figure 1, where it has nonzero entries if the drug is effective for the corresponding nodes (for patients) or is the node itself (for drug nodes). The goal is to predict the values of those zero entries (for patient nodes, those are the entries indicating whether this drug will be effective or not for them; for drug nodes, those are the entries indicating whether this drug would be similar to them in real-world clinical usage). If we concatenate all effectiveness vectors for the m drugs, we can form a drug effectiveness matrix $\boldsymbol{Y}=[\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_m]$. Then we adopted a label propagation procedure to spread the label information in $\boldsymbol{Y}$ for the whole graph. Over this heterogeneous graph, patients propagate their known effective treatments to other patients based on the patient similarity analytics, and drugs propagate their target effective patients to other drugs based on the drug similarity analytics simultaneously to derive the relevance between nodes until achieving a steady state. After label propagation, possibilistic label (i.e., the possibility when a drug is effective for a patient) matrix $\boldsymbol{F}$ can be obtained by a formula $\boldsymbol{F}=(1-\mu)(\mathbf{I}-\mu\boldsymbol{W})^{-1}\boldsymbol{Y}$ (for details please refer to Wang and Zhang[23]). In this formula, $\boldsymbol{W}$ is a normalized form of the similarity matrix $\boldsymbol{A}$, and $0<\mu<1$ is a parameter that determine the influence of a node's neighbors relative to its provided label.

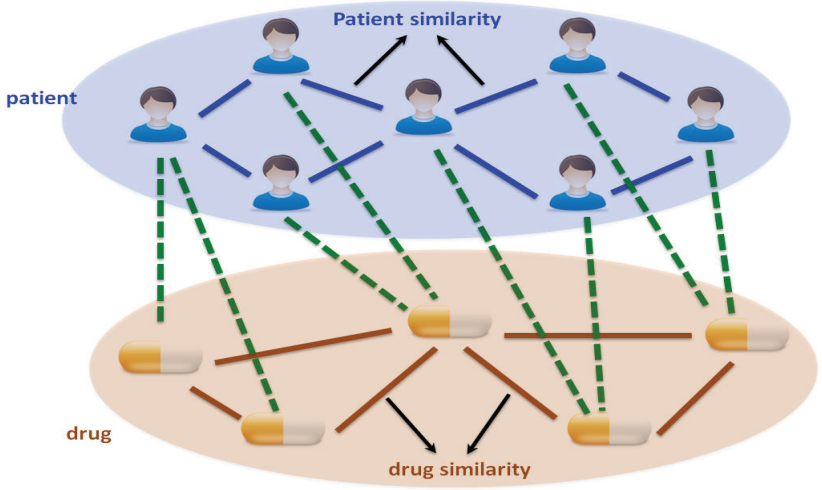


Figure 1. Illustration of the proposed heterogeneous label propagation method. The heterogeneous graph constructed with patients and drugs, where patient is one domain and drug is another domain. There are three types of relationships encoded in this graph: patient similarities, which are the blue edges; drug similarities, which are the yellow edges; patient-drug prior associations, which are the green dashed edges.

Results

In this section we present experimental evaluation results of the proposed heterogeneous label propagation method on a treatment recommendation task for individual patients.

Data Description. Our real-world dataset contains 3-year longitudinal EHR of 110,157 patients. We selected *hypercholesterolemia* as our target disease for conducting experimental evaluations. There are 8 cholesterol-lowering drugs and 273,525 Low-Density Lipoprotein (LDL) lab-test records in the dataset. A patient, whose LDL level is below 130 mg/dL, is considered to be "well-controlled". To define an effective drug for a patient, we selected the patients who take only one cholesterol-lowering drug within a 60-day treatment window and remain "well-controlled" for at least two consecutive lab assessments. We obtained 1219 distinct patients and 4 statin cholesterol-lowering drugs (i.e., *Atorvastatin* effectively treats 97 patients, *Lovastatin* effectively treats 221 patients, *Pravastatin* effectively treats 24 patients, and *Simvastatin* effectively treats 877 patients). The drug similarities from chemical structures and drug targets were calculated respectively. The patient similarities were calculated based on the ICD9 diagnosis codes within the 90-day patient assessment window prior to the first day a patient takes a drug within the 60-day treatment window. Then we constructed a heterogeneous graph based on our proposed method.

For illustration, Figure 2 depicts the definition of an effective drug for a given patient and assessment of patient diagnosis condition prior to treatments.

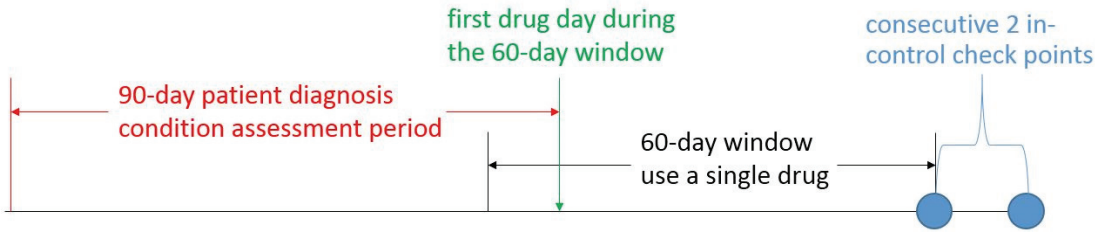


Figure 2. Assessments of patient diagnosis condition prior to treatments and definition of the effective drug for a single patient over time. Blue circles represent "well-controlled" LDL assessments (LDL < 130 mg/dL).

Method Comparison. We used a 10-fold cross-validation scheme to evaluate treatment recommendation algorithms. To obtain robust results, we performed 50 independent cross-validation runs, in each of which a different random partition of the dataset to 10 parts was used. In our comparisons, we considered three treatment recommendation methods: (1) Label propagation using only patient information. The method propagates known effective treatments of training patients to testing patients based on the patient similarity analytics without considering drug information. (2) Heterogeneous label propagation using both patient and drug chemical structure information. The method propagates known effective treatments of training patients to the whole heterogeneous graph which is proposed in the methodology section. The drug similarity is calculated based on drugs' chemical structures. (3) Heterogeneous label propagation using both patient and drug target information. The method propagates known effective treatments of training patients to the whole heterogeneous graph and the drug similarity is calculated based on drugs' protein targets. Figure 3 shows the averaged ROC curves of 50 runs of the cross-validation for different methods based on the experiment.

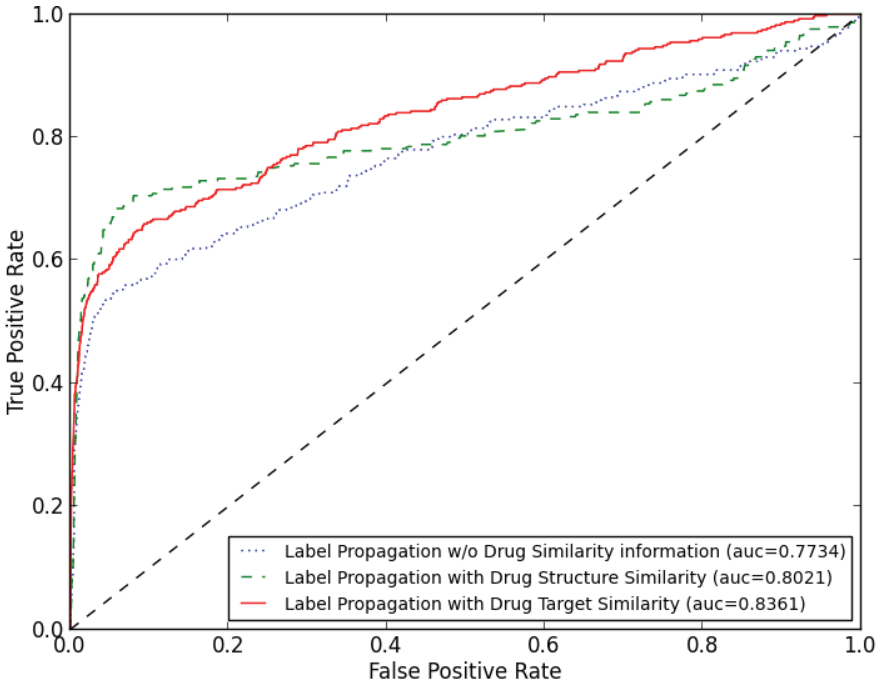


Figure 3. The averaged ROC comparison of three treatment recommendation strategies. Methods are sorted in legend of the figure according to their AUC score.

Figure 3 shows that label propagation algorithms are capable at treatment recommendation tasks. Without using any drug information, the label propagation algorithm obtains an averaged AUC score of 0.7734. When combining drug chemical structure or drug target information, heterogeneous label propagation algorithms obtain averaged AUC scores of 0.8021 or 0.8361 respectively. Analysis of the results revealed that rarely used treatments in the EHR data (e.g., *Pravastatin* only has 24 effective cases in the data, but it is very similar to *Lovastatin* from both structure and

target perspectives) benefit from drug similarity analytics, thus the overall AUC scores were improved. Another observation is that heterogeneous label propagation using drug target similarity achieved a higher AUC score (0.8361) than the one using drug chemical structure similarity (0.8021). The results indicate that choosing an appropriate drug similarity measurement for the dataset will improve the performance of the heterogeneous label propagation. For example, *Lovastatin* is used to lower LDL by less than 30%, *Simvastatin* is used to lower LDL by 30% or more and treat the patients have heart disease and/or diabetes in the clinical settings[24]. *Lovastatin* and *Simvastatin* have very similar chemical structures, thus chemical structure similarity may not distinguish them well. Instead, *Lovastatin* and *Simvastatin* have different drug target sets (i.e., *Lovastatin* targets proteins *3-hydroxy-3-methylglutaryl-coenzyme A reductase*, *Integrin alpha-L*, and *Histone deacetylase 2*; *Simvastatin* targets proteins *3-hydroxy-3-methylglutaryl-coenzyme A reductase*, and *Integrin beta-2*), thus in this study drug target similarity may serve as a better similarity metric to recommend personalized treatments to patients.

Conclusion

We have proposed a heterogeneous label propagation method to support personalized medicine by leveraging patient similarity and drug similarity analytics. Experimental evaluation results on a real-world EHR dataset demonstrate the effectiveness of the proposed method and suggest that the combination of appropriate patient similarity and drug similarity analytics can help identify which drug is likely to be effective for a given patient. In future work we plan to apply the method to more drugs and more diseases, and explore more sophisticated drug and patient similarity measures.

References

1. Meyer UA. Pharmacogenetics - five decades of therapeutic lessons from genetic diversity. Nat. Rev. Genet. 2004;5:669-675.
2. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB. Bioinformatics challenges for personalized medicine. Bioinformatics 2011;27(13):1741-1748.
3. Neuvirth H, Ozery-Flato M, Hu J, Laserson J, Kohn MS, Ebadollahi S, Rosen-Zvi M. Toward personalized care management of patients at risk: the diabetes case study. In Proceedings of ACM international conference on knowledge discovery and data mining 2011:395-403.
4. Liu L, Tang J, Cheng Y, Agrawal A, Liao WK, Choudhary A. Mining diabetes complication and treatment patterns for clinical decision support. In Proceedings of ACM international conference on information and knowledge management 2013.
5. Rosen-Zvi M, Altmann A, Prosperi M, Aharoni E, Neuvirth H, Sonnerborg A, Schülter E, Struck D, Peres Y, Incardona F, Kaiser R, Zazzi M, Lengauer T. Selecting anti-HIV therapies based on a variety of genomic and clinical factors. Bioinformatics 2008;24(13):i399-i406.
6. 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.
7. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol. 2011;7:496.
8. Li J, Lu Z. A new method for computational drug repositioning using drug pairwise similarity. In Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine 2012.
9. Zhang P, Agarwal P, Obradovic Z. Computational drug repositioning by ranking and integrating multiple data sources. In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2013:579-594.
10. Lounkine E, Keiser M et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature 2012;486:361-367.
11. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Brief Bioinform. 2013.
12. Gottlieb A, Stein GY, Oron Y, Ruppin E, Sharan R. INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol. 2012;8:592.
13. Vilar S, Harpaz R, Uriarte E, Santana L, Rabadan R, Friedman C. Drug-drug interaction through molecular structure similarity analysis. J Am Med Inform Assoc. 2012;19(6):1066-1074.
14. Sun J, Wang F, Hu J, Edabollahi S. Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 2012;14(1):16-24.
15. Wang F, Hu J, Sun J. Medical prognosis based on patient similarity and expert feedback. In Proceedings of International Conference on Pattern Recognition 2012:1799-1802.
16. Chawla NV, Davis DA. Bringing big data to personalized healthcare: a patient-centered framework. J Gen Intern Med. 2013.
17. Syed Z, Guttag JV. Unsupervised similarity-based risk stratification for cardiovascular events using long-term time-series data. Journal of Machine Learning Research 2011:999-1024.
18. Roque FS, Jensen PB et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011;7(8):e1002141.
19. Huang Z, Dong W, Duan H, Li H. Similarity measure between patient traces for clinical pathway analysis: problem, method, and applications. IEEE Journal of Biomedical and Health Informatics 2013.
20. Wishart DS, Knox C et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(Database issue):D668-D672.
21. Smith TF, Waterman MS, Burks C. The statistical distribution of nucleic acid similarities. Nucleic Acids Res. 1985;13(2):645-656.
22. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. J Am Med Inform Assoc. 2013;20(5):954-961.
23. Wang F, Zhang C. Label propagation through linear neighborhoods. In Proceedings of International Conference on Machine Learning 2006:985-992.
24. Evaluating statin drugs to treat high cholesterol and heart disease: comparing effectiveness, safety, and price. Best Buy Drugs (Consumer Reports):9.

The ENCODE ChIP-Seq Significance Tool: Enabling the Study of Disease Mechanisms Through the Use of Public Data

Raymond K Auerbach, Bin Chen, Atul J Butte
Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, 94305

Summary:
High-throughput genomic methods have released a treasure trove of data that, when integrated, can lead to a better understanding of disease. However, simple yet powerful tools that enable biomedical researchers to quickly mine these resources are still needed. We present the ENCODE ChIP-Seq Significance Tool, a web application available at http://encodeqt.stanford.edu that mines ENCODE ChIP-Seq data to identify enriched transcription factors given a list of genes or transcripts.

Background:
High-throughput methods enabled by improvements in engineering, computing, and the sciences have resulted in a plethora of data being produced. In the biological sciences, data from high-throughput, next-generation sequencing assays can provide insights into how different mechanisms such as those related to epigenetics, transcription, and DNA-binding work in concert to regulate biological processes. Understanding how transcription is regulated can provide insights into why we exhibit certain characteristics and even help us understand the mechanisms that may cause a particular disease. International consortia have been generating many different types of data that can be integrated with this goal in mind. The ENCODE Consortium alone has spent over $200 million to generate several data sets exploring transcription factor binding, gene expression, and DNA accessibility across 357 different human cell types.[1] In particular, ENCODE ChIP-Seq data that identifies transcription factor binding sites on a genome-wide scale is useful for relating gene expression to gene regulation, but with such vast amounts of data, a need for user-friendly tools for biomedical scientists that leverage these data becomes paramount. We present the ENCODE ChIP-Seq Significance Tool, a flexible web-based application to mine these data (http://encodeqt.stanford.edu).

Methods:
We obtained peak regions from 708 ChIP-Seq experiments encompassing 220 transcription factors and cellular treatment combinations across 91 cell types from the ENCODE Consortium download site. We then intersected the position of each peak apex against the start and end positions of each gene, identifying the closest peak to the transcription start site (TSS) and transcription termination site (TTS). These values were recorded in a database for each gene/transcription factor/cell line combination. Our flexible web application leverages this database and allows users to fine-tune parameters such as which cell lines to consider, which window size to use, and which genes to use as a background set. The latter feature allows our tool to be used for both microarray and sequencing-based experiments. Given a list of gene or transcript IDs, a hypergeometric test is run for each transcription factor/cellular treatment combination to identify enriched transcription factors. After Benjamini-Hochberg multiple hypothesis correction, transcription factors are ranked by q-value and associated data. Genes and transcripts that intersect ChIP-Seq peaks for each factor can also be retrieved.

Results/Discussion:
By using our tool to leverage public ENCODE ChIP-Seq data, researchers can go beyond gene signatures and begin to explore the underlying mechanisms that cause disease. Identifying which transcription factors are regulating target genes of interest is a major step towards understanding the relationship between these genes in a network context, identifying potential drug targets, and ultimately designing new therapeutics. This work was published in *Bioinformatics* and has been featured in several blogs (Getting Genetics Done, StatsBlogs, and OpenHelix) as a useful tool for the scientific community.[2] New features implemented since publication will also be presented.

References:
[1]ENCODE Project Consortium et al (2013). An integrated encyclopedia of DNA elements in the human genome. *Nature*. 489:57-74.

[2]Auerbach RK, Chen B, Butte AJ (2013). Relating Genes to Function: Identifying Enriched Transcription Factors using the ENCODE ChIP-Seq Significance Tool. *Bioinformatics*. 29(15):1922-1924.

Using SemRep and a medication indication resource to extract treatment relations from clinical notes

Cosmin A. Bejan, PhD[1], Wei-Qi Wei, MD, PhD[1], Joshua C. Denny, MD, MS[1,2]

[1]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN;
[2]Department of Medicine, Vanderbilt University, Nashville, TN

Abstract: The goal of this study is to evaluate the contribution of SemRep and a medication indication (MEDI) resource to the task of extracting treatment relations from clinical notes. Although in many cases these relations link medications to diseases, there exist other types of treatment relations such as procedure-disease, procedure-patient, etc. Our preliminary results show that MEDI has a positive impact to this task when combined with SemRep.

Introduction and Background: Providers often record the reasons (i.e., the indications) for therapeutic interventions in their clinical notes. Our purpose was to investigate the impact of a medication indication resource on an existing relation extraction system for discovering treatment relations in clinical text. As a medication indication resource we selected MEDI,[1] a large database of medication-indication pairs, and as an extraction system we used SemRep,[2] a publicly-available and widely-used system successfully applied to literature data sets. Our ultimate goals are to create an automatic extraction system that will improve systems like SemRep for identifying treatment relations, and to expand MEDI with new medication-indication pairs. The ability of accurately extracting treatment relations could enable a more comprehensive understanding on a patient's treatment course, improve adverse reaction detection, discover off-label drug uses, and allow public health surveillance for common diseases.

Methods: In our study, we used a set 6864 discharge summaries from the Vanderbilt Synthetic Derivative, a de-identified version of the Vanderbilt electronic medical record. First, we processed the reports with SemRep (v1.5). For each sentence, the system extracted all corresponding UMLS concepts and the treatment relations between them. Next, we analyzed the concepts identified by SemRep (regardless of whether SemRep found a relationship between them) to identify possible medication-indication pairs from MEDI that co-occurred within one sentence. To evaluate how accurately SemRep and MEDI discovered treatment relations, two reviewers annotated the sentences in which both resources identified at least one relation. The annotation process consisted of manually linking pairs of UMLS concepts that represent treatment relations, limited to the identified UMLS concepts and blinded to the algorithms' results. The reviewers performed double annotations on 75% of the data. The inter-annotator agreement reached a Cohen's kappa value of 0.86 and the disagreements were adjudicated by an experienced clinical expert.

Results and Conclusion: After data processing, 943306 UMLS concepts and 3386 treatment relations were identified by SemRep, and 1590 UMLS concept pairs were matched to medication-indication concept pairs from MEDI (Figure 1). The small overlap of relations shown in Figure 1 is mainly due to the existence of non-medication relations (e.g., procedure-disease) that MEDI is not able to capture. In our evaluation, we compared the manual annotations (393 treatment relations and 4177 non-treatment relations) against the SemRep and MEDI relations. We also considered two simple ensemble methods which combine the predictions of the two resources – the union and intersection of SemRep and MEDI. As seen in Table 1, both MEDI and SemRep performed similarly with reasonable precision, despite having only 150 (3.1%) relations commonly identified over the entire dataset (Figure 1). The F-measure obtained by MEDI was slightly better than the SemRep, despite not using any linguistic information. The best results, in terms of F-measure, were achieved by the union of MEDI or SemRep results. The improvements of this method over both SemRep and MEDI stem primarily from significant gains in recall and comparatively smaller loss in precision. Not surprisingly, the more restrictive method (i.e., MEDI and SemRep) achieved high precision at the cost of large drops in recall. Further investigation is needed for research and clinical uses of such data.

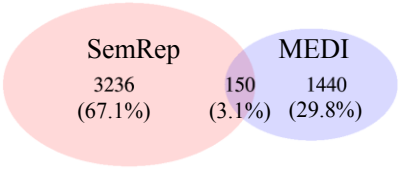


Figure 1 The connection between the relations identified by SemRep and MEDI.

Configuration	Precision	Recall	F-measure
MEDI	79.85	54.45	64.75
SemRep	76.70	54.45	63.69
MEDI and SemRep	93.66	33.84	49.72
MEDI or SemRep	72.84	75.06	73.93

Table 1 Results for treatment relation extraction.

References

1. Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. J Am Med Inform Assoc. 2013 Sep 1;20(5):954-61.
2. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.

An integrated framework for the pharmacogenomic characterization of oncological drug response to enable precision medicine

Krithika Bhuvaneshwar, MS[1], Michael Harris, MA[1], Thanemozhi Natarajan, PhD[1], Laura Sheahan, PhD[1], John Deeken, MD[2], Subha Madhavan, PhD[1]
[1]Innovation Center for Biomedical Informatics, Georgetown University Medical Center, Washington DC; [2]Inova Translational Medicine Institute, Virginia

Summary

Genetic variations in genes involved in absorption, distribution, metabolism, and excretion (ADME) can alter drug response in some patients leading to adverse reactions such as toxicity or resistance. We have developed a framework to identify variants that are significantly associated with drug response in an effort to enhance precision medicine, and present here two applications– response to gemcitabine and doxorubicin in cancer patients.

Introduction and Background

Genes involved in absorption, distribution, metabolism, and excretion (ADME) are known to influence the pharmacological activity of drugs. These drugs have various adverse effects, which can be life-threatening. Earlier studies based on cell lines and patient-control populations have demonstrated that inter-individual variations in germline DNA can impact cellular response to oncology drugs. The Affymetrix Targeted Human Drug Metabolizing Enzymes and Transporters (DMET) 1.0 chip enables genotyping of 1,256 known variants in 170 ADME genes. These known variants in ADME genes have low predictive power for many complex diseases including cancer. The use of statistical and probabilistic methods in conjunction with novel informatics, data integration and systems biology approaches are helping to identify clinically relevant variants. Identification and validation of such variants that have a functional/regulatory impact can help understand the complex pharmacologic pathways of anti-cancer drugs and ultimately help develop clinical decision support tools that predict toxicity and efficacy of chemotherapeutic agents and enable precision medicine in cancer care.

Methods

We analyzed genotype data from NCI-60 cell lines using the DMET chip combined with drug sensitivity data, denoted by its GI50 values on two oncology drugs – gemcitabine (*gem*) and doxorubicin (*dox*). Fisher's exact test (FET) and Probabilistic Network Analysis (PNA) were conducted to identify variants significantly associated with drug response. The significant genes from each analysis were validated against literature, and pathway analysis was performed to determine their potential role in drug metabolism. We performed an integrated analysis of gene expression, SNP and drug response data for the NCI-60 cell lines using t-tests and linear regression models to see how SNPs influence gene expression levels via cis- or trans-regulatory effects.

Results and Discussion

We identified several variants strongly associated with *dox* such as ABCC2 (rs3740066, FET p-value 0.0333, coding-syn) and ABCC6 (rs2238472, FET p=0.0105, missense) that mediate the transport of glutathione and are known to cause chemo resistance. *Dox*, a known substrate for ABCC2 (MRP2), is regulated by the ERKS - MAPK3 (ERK1) and MAPK1 (ERK2), and in turn activates these genes in cancer cells. This leads to up-regulation of anti-apoptotic and pro-survival genes along with other ABC transporters that confer *dox* resistance in cancer cells. Thus abnormal activation of the Raf/MEK/ERK signaling, a frequent event in many cancers, can induce chemoresistance to many drugs including *dox*.

Genetic variants in the ABC transporters ABCC1 (rs8187858, FET p=0.001, cds-synon) and ABCC4 (rs4148551, FET p=0.01, utr-3), CHST3 (rs4148943, FET p=0.002, utr-3), and PPARD (rs3798343, FET p=0.049, intron) were found to be significantly associated with *gem* response. We identified novel variants in the CHST genes that were significantly associated with *gem* response. Members of the CHST family are known to mediate inflammation, immunity, angiogenesis, and extracellular matrix reorganization, and oncogenic HRAS signaling.

Conclusion

This integrative methodology found important variants that should be confirmed. Further, our methodology could be applied to other oncology drugs to help gain insights into the genomic causes of drug response. This approach and the results could be used for the development of clinical decision support tools and means to personalize anticancer drug therapy.

Searching for master regulators of disease-related gene expression profiles by network analysis of the LINCS library of transcriptional signatures of cellular perturbations

[1]Mario Medvedovic, PhD, [1]Jing Chen, PhD, [1]Mukta Phatak, PhD, [1]Siva Sivaganesan PhD, [2]John Reichard PhD, [1]Wen Niu, MS, [1]Vineet Joshi, MS

Affiliation: [1]University of Cincinnati, [2]TERA; Location: Cincinnati, Ohio

Summary: We have developed an analytical framework for studying regulatory networks disrupted by the disease. A *regulatory network model* defines a set of regulatory proteins, transcription factors and their interactions that lead to disease-related gene expression changes. Understanding regulatory networks governing gene expression changes helps us understand causes of the disease, and design effective treatment and prevention strategies. Our framework utilizes novel computational algorithms and a large-scale library of transcription factor binding and regulatory protein perturbation signatures derived from more than one million genome-wide profiles of gene expression levels generated by the LINCS project (http://LincsProject.org). The effectiveness of the methodology is demonstrated in the analysis of transcriptional signatures of driver mutations in TCGA data and the transcriptional signatures of estrogenic compounds. Computational tools and data needed to perform NetLincs analysis can be downloaded from iLINCS portal (http://LincsGenomics.org).

Background: Understanding the etiology of human diseases at the molecular level is necessary for designing effective treatments and prevention strategies. Studying genome wide gene expression profiles associated with a disease has been an effective way to shed light on disease-related biological processes. Regulatory models define causative regulatory events that lead to a disease-related transcriptional profile. Since gene expression measurements define mostly consequences of such regulation, it is usually difficult, if not impossible, to infer a regulatory model from gene expression data alone.

Methods: To identify regulatory proteins whose activity can explain the Disease-Related Gene Expression Profile (DRGEP) of interest, we first perform concordance analysis between the DRGEP and transcriptional activity signatures of more than 4,000 regulatory proteins and transcription factors in different cell lines at different time-points. Regulatory activity signatures were derived from LINCS perturbation signatures (http://LincsProject.org). Each signatures consists of genome-wide (ie for each gene) differential expression levels and associate p-values calculated by comparing gene expression levels after a perturbation of a regulatory protein to the control. The DRGEP of interest is first correlated with each perturbation signatures. Resulting correlations represent regulatory activity scores of the perturbed protein in producing the DRGEP. Protein-protein interactions between perturbed proteins documented in the STRING database were used to construct global regulatory network. Regulatory activity levels of different perturbed proteins were calculated by graph diffusion kernel based network analyses of activity scores. Statistically significant regulatory proteins are established by comparing their activity levels to the empirical null-distribution established by randomly permuting activity scores among the regulatory proteins.

Results: We analyzed breast cancer mutational signatures for genes established to be mutated more commonly than expected by chance by the TCGA consortium paper. For each such gene, a DRGEP was constructed by comparing expression profiles of samples with deleterious mutation in a given gene with all other samples. Ability of our methodological framework to predict the mutated genes was assessed by ROC curves. We show that our methodology produces a strong signal capable of predicting the mutation. We also show that the network modeling step dramatically improves the predictive ability of activity measures when compared to simply using level of correlation with the perturbation signature to predict the mutated regulator. Similar results were obtained in searching for evidence of ERα activity after treatment of the MCF-7 cell line with a series of estrogenic chemicals. Finally, we perform the validation of the strategy by analyzing all perturbation signatures in one cell line (MCF-7) using perturbation signatures in another, not closely related cell line (PC3). The area under the curve of this analysis was 0.93 indicating excellent reproducibility of the results in this setting.

Discussion: Our analysis indicates a strong potential for this analytical strategy in which a large-scale library of gene perturbation signatures is used to identify master regulators of disease related transcriptional signatures. In addition to LINCS data, significant portion of the genomics profiles in the major public genomics data repositories directly measure transcriptional effects caused by specific perturbations of the global regulatory network. The perturbations can be in the form of siRNA-based expression knockdowns, genetic manipulations or over-expression experiments. These grossly underutilized data may be turned into a valuable resource by incorporating them within an analytical platform as described here. The limits of the strategy that uses signatures generated in in-vitro systems are still to be established. Our analysis indicates that at least in some disease-related signatures, this strategy can be very effective.

Preliminary Classification of Cancer Sites Using Machine Learning and Somatic Mutations from the COSMIC Database

Yukun Chen[1], Yaoyi Chen[1], Michael A. Kochen[1], Zhongming Zhao[1,2,3], Hua Xu[4,1]

[1]Department of Biomedical Informatics, [2]Department of Cancer Biology, and [3]Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN, USA
[4]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Summary
This is a preliminary study to explore the somatic mutation patterns among different cancers using a machine learning algorithm and COSMIC database. We performed a multi-class classification experiment over 16 tumor sites using the gene symbols and somatic mutation descriptions as predictors for 3502 subjects.

Background:
Recent studies suggest that different cancers might have shared some common driver genes through their somatic mutations, but the sharing patterns remain largely unclear to us. We aimed to explore the mutation patterns among different types of cancer by machine learning algorithm using the Catalogue Of Somatic Mutations In Cancer (COSMIC) database, which is a curated, comprehensive somatic mutation database for human cancer.

Methods:
We built a support vector machine (SVM) based predictive model for resolving multiple-cancer classification. We extracted 3502 subjects who had at least 10 mutations in the COSMIC database. Two feature sets were tested in this study. The first set used gene symbols, and the second set included both gene symbols and their associated mutation descriptions. There were a total of 16,689 unique gene symbols and 15 descriptions of mutations (complex, deletion, insertion, etc.) from the dataset. The size of unique features that combined gene symbols and mutation descriptions was 62,256 after we removed those features that only appeared once in the dataset. A one-versus-all scheme was implemented to solve the multi-class classification task. The ten-fold cross-validation was performed to evaluate the predictive models.

Results:
SVM achieved an accuracy of 0.59 using gene symbols only as the features. The performance was improved to 0.65 when adding the information of mutation descriptions. Table 1 summarizes the performance over 16 primary cancer sites. Lung, kidney, and pancreas are the primary sites where SVM could achieve greater than 0.70 in Fmeasure using gene symbols as features. Using mutation descriptions as well, the performances on lung and kidney were improved from 0.77 to 0.78 and 0.73 to 0.82, in Fmeasure, respectively; however, the performance on pancreas data was decreased from 0.73 to 0.69. The performance on most other primary sites was improved as well, with the exception of urinary tract. We could not identify autonomic ganglia using both types of feature sets.

Discussion:
SVM could accurately identify some primary sites of cancer using information of Gene and associated mutation up to 0.82 in Fmeasure (e.g. Kidney cancer). However, the model was not explicitly interpretable because the model was in the dimension of over 60,000. In the future, we would run machine learning experiment using the features reduced to, for example, the set of most frequently mutated genes at different primary sites.

Table 1. Result of multiple-primary-site classification task

	Gene symbols			**Gene symbols and mutation descriptions**		
	Accuracy = 0.59			**Accuracy = 0.65**		
Primary site (n, %)	**Precision**	**Recall**	**Fmeasure**	**Precision**	**Recall**	**Fmeasure**
Lung (721, 20.6%)	0.82	0.72	0.77	0.85	0.73	0.78
Kidney (351, 10.0%)	0.79	0.68	0.73	0.85	0.79	0.82
Pancreas (200, 5.7%)	0.73	0.74	0.73	0.68	0.70	0.69
Large intestine (286, 8.2%)	0.53	0.85	0.65	0.58	0.86	0.69
Endometrium (204, 5.9%)	0.62	0.64	0.63	0.67	0.75	0.71
Liver (67, 2.0%)	0.60	0.63	0.61	0.85	0.51	0.64
Stomach (25, 0.8%)	0.71	0.48	0.57	0.77	0.40	0.53
Haematopoietic and lymphoid tissue (246, 7.1%)	0.53	0.58	0.56	0.54	0.67	0.60
Skin (90, 2.7%)	0.46	0.50	0.48	0.82	0.50	0.62
Central nervous system (221, 6.4%)	0.46	0.47	0.47	0.48	0.46	0.47
Breast (308, 8,8%)	0.41	0.50	0.45	0.47	0.56	0.51
Ovary (338, 9.7%)	0.43	0.43	0.43	0.56	0.61	0.58
Prostate (268, 7.7%)	0.46	0.38	0.42	0.51	0.46	0.49
Urinary tract (51, 1.6%)	0.57	0.31	0.41	0.71	0.24	0.35
Upper aerodigestive tract (104, 3.1%)	0.45	0.24	0.31	0.64	0.53	0.58
Autonomic ganglia (22, 0.6%)	0	0	0	0	0	0

Creating Scalable Research Infrastructure to Enable Translational Science: the Synthetic Derivative

James Cowan, BS, Melissa Basford, MBA, Xiaoming Wang, MS, Susan Osgood, BS, Paul Harris, PhD, Joshua C. Denny, MD, MS
Vanderbilt University School of Medicine, Nashville, TN

Summary: The proliferation of electronic health records (EHRs) has created opportunities for re-using EHR data for biomedical research. Efforts to link EHRs to DNA biobanks have demonstrated utility in using EHR data for genomic research. We describe our 3-year experience with one such effort at Vanderbilt University Medical Center.

System Description: Vanderbilt University Medical Center (VUMC) has EHR data for over 2 million patients. Data availability to researchers is made via our de-identified database and user interface tools, collectively called the Synthetic Derivative (SD), which is linked to BioVU, a de-identified DNA biobank including >170,000 samples. De-identified, unstructured clinical text is stored alongside structured data such as lab results, diagnosis and procedure codes, and physician orders. Various data transformations and natural language processing (NLP) tools are applied, enriching the data to support use cases across the entire clinical and translational research spectrum.

Three tiers of user access are supported. The Record Counter web application (RC) allows querying of all data sources in the SD database, but limits results to aggregated counts grouped by demographics. A web application eponymously called the Synthetic Derivative shares a query interface with the RC, but also supports review of full de-identified records. Direct access to underlying database tables supports complex use cases where researchers with more advanced database skills need to interact with the data directly.

Evaluation: To evaluate usage of the SD, surveys were sent to three groups of people: users of the RC, users of the SD, and users who access both systems. Survey results determined the types of users (faculty, staff, etc.), their objective(s) (hypothesis generation, support of clinical studies, etc.), significant results (publications, grant awards) and their perception of the SD's utility. 651 SD/RC system users (335 RC, 255 SD, 61 SD/RC) were invited to participate in an anonymous REDCap survey. 213 responses were received (109 RC, 81 SD, 23 SD/RC), for a total response rate of 33% and individual tool response rates of 33%, 32%, and 38%, respectively.

Results: Among all survey respondents, Faculty represented 55%, Staff 21%, Post-Doc/Resident/Fellows 14% and Students 10%. To date, 122 publications have used data from the SD. 31 grant awards have either used SD data or the tools themselves for research support. Users who would recommend the tool(s) to others doing similar work were 97 (89%) RC, 72 (89%) SD , and 23 (100%) SD/RC. A retrospective review of user queries suggests all major domains of disease (assessed by ICD-9-CM chapters) are represented, led by cancers and endocrine diseases. Survey respondents self-reported diverse use cases when asked how the RC and SD applications were utilized to support individual projects:

RC Usage		**SD Usage**	
Clinical Study Feasibility	57 (43%)	BioVU Genomic Study	51 (49%)
Hypothesis Generation	51 (39%)	Hypothesis Generation	40 (38%)
Preliminary Data for Grant Submissions	43 (33%)	Clinical Study	32 (31%)
BioVU Genomic Study Feasibility	40 (30%)	Preliminary Data for Grant Submissions	28 (27%)
Clinical Trial Feasibility	21 (16%)	Non-BioVU Genomic/Proteomic Study	10 (10%)
Quality Improvement Projects	18 (14%)	Quality Improvement Projects	7 (7%)
Other	24 (28%)	Other	8 (8%)

Discussion: Repurposed EHR data can be a valuable resource for clinical and genomic discovery. We found that a broad range of users investigated both clinical and genomic questions using the SD, and that it has rapidly become an important tool for supporting publications and grants.

Phenome Wide Association Studies demonstrating pleiotropy of a genetic variant within *FTO* with and without adjustment for BMI

Rob Cronin, MD[1], Julie Field, PhD[1], Lisa Bastarache, PhD[1], Dana Crawford, PhD[1], Josh Denny, MD, MS[1]; [1]Vanderbilt University, Nashville, TN

Abstract/Summary

Phenome wide association studies (PheWAS) have demonstrated utility in validating genetic associations derived from traditional genome-wide association studies (GWAS) as well as identifying novel genetic associations. Here we provide an example of EHR-based PheWAS to explore pleiotropy of FTO rs8050136 associated with obesity.

Introduction and Background

A complement to the genome-wide association study (GWAS), Phenome wide association studies (PheWAS) enables both the validation of genotype-phenotype associations identified by GWAS and the generation of new hypotheses, identifying potentially novel associations in need of further investigation as well as putative instances of genetic pleiotropy.

Methods

A population of 13,711 individuals of European Ancestry contained within the BioVU DNA databank was genotyped using the Illumina Infinium HumanExome BeadChip. Genotyping quality was evaluated using Single Nucleotide Polymorphism (SNP) call rates and concordance rates with HapMap controls; SNPs with <90% call rate or <98% concordance were excluded. Rs8050136, which had a call rate of > 99.9%. All individuals had identity by descent estimates greater than 0.25. PheWAS analysis was carried out with methods described previously using the PheWAS R package[1]. We performed pairwise analysis of each case and control group for the analyzed SNP using an additive model with logistic regression and adjusted for age, sex, and the first three principal components. We performed the PheWAS with adjustment for average Body Mass Index from our EHR.

Results and Discussion

We noted the well described associations of obesity ($p=1.36x10^{-6}$) in our analysis when analyzed without adjustment for BMI. Three obesity-related diseases also trended toward significance: type 2 diabetes (T2D; $p = 5.3x10^{-5}$) and obstructive sleep apnea (OSA; $p = 4.6x10^{-3}$). The associations with obesity and OSA were largely attenuated by adjustment for BMI (obesity: p=0.055; OSA: p=0.689). Adjusting for BMI significantly attenuated the association with T2D (p=0.015). However, some associations significant at an unadjusted p of 0.01 were unchanged at BMI adjustment including: benign mammary dysplasias, joint effusions, obstruction of bile duct, streptococcus and staphylococcus infections, disease of the pulp and periapical tissues (Figure 1). More investigation is needed to determine the possible pleiotropic nature of these associations. This analysis highlights the potential power of the PheWAS method to identify pleiotropic genes as putative contributors to multiple comorbid diseases and traits.

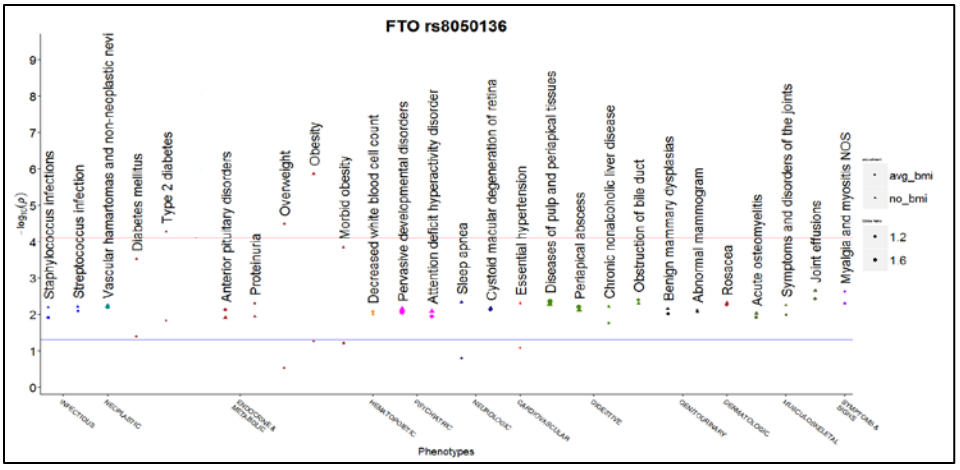


Figure 1: PheWAS analysis of *FTO* without and with average BMI adjustment. Phenotypes with p < 0.01 are labeled.

References

1 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* Published Online First: 24 November 2013. doi:10.1038/nbt.2749

The Pharmacogenomic Guideline Repository: A Resource of Structured Guidelines to Facilitate Clinical Implementation

Robert R. Freimuth, PhD, Qian Zhu, PhD, and Christopher G. Chute, MD, DrPH

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract

The Pharmacogenomic Guideline Repository (PGR) contains pharmacogenomic (PGx) guidelines in a structured and coded format. This is a first step toward providing machine-readable representations that could facilitate the adoption of PGx guidelines by simplifying their integration into existing clinical infrastructure, including electronic medical record (EMR) systems and Clinical Decision Support (CDS) engines.

Introduction and Background

The number of "clinically-actionable" genetic variants is increasing rapidly due to advances in PGx knowledge[1]. To help clinicians make decisions about prescriptions based on a patient's genotype, PGx guidelines are being published[2]. While these guidelines are tremendously valuable to individual practitioners, they are not yet available in a computable form that facilitates their adoption at an institutional level, and the effort required to adopt, integrate, and maintain these guidelines within clinical infrastructure remains significant. In particular, the guidelines contain unstructured text, lack standard identifiers, and are not available in an easily parsable format. The Clinical Decision Support Consortium (CDSC) proposed a layered framework to translate unstructured text (Level 1) into executable code (Level 4)[3]. We applied this approach to published PGx guidelines (Level 1) to create structured forms (Levels 2 and 3) as a step toward executable representations (Level 4).

Methods

Published PGx guidelines were used to inform the development of an information model, which provided a structure for guideline content. The model was extended to include references to standard identifiers for genes (HGNC) and genetic variants (dbSNP), and standard terminologies were used to annotate concepts for drug ingredients (RxNorm) and molecular phenotypes (LOINC). The use of standards will facilitate integration with drug order entry and CDS systems as well as provide a mechanism to link the guidelines to public knowledgebases. A versioning scheme was added to support guideline updates. The information model was expressed in UML and converted to an XSD schema. Pharmacogenomic guidelines were represented in XML and validated against the XSD schema.

The PGR web service was built using the Grails web application framework and an Apache CouchDB database. The system includes an administrative browser-based interface to upload structured guidelines and a REST programming interface. Several query functions are available.

Results and Discussion

Evaluation of the information model using XML representations of published PGx guidelines demonstrated that the model successfully structured the core content and semantics of the original guidelines, but highlighted the need for standardized allele nomenclature and molecular definitions. Ongoing efforts include modeling the supporting information within PGx guidelines using tooling from the CDSC and Health eDecisions.

Structuring PGx guidelines and incorporating standard identifiers will streamline the maintenance of guidelines that have been implemented within clinical systems. In addition, it will reduce the potential for human error and help to ensure more consistent interpretations of drug dosing guidelines across institutions. This resource is a first step towards creating sharable, executable representations for PGx guidelines.

References

1. Table of Pharmacogenomic Biomarkers in Drug Labels [2013 October 10]. Available from: http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm
2. PharmGKB CPIC Gene Drug Pairs [2013 October 10]. Available from: http://www.pharmgkb.org/page/cpicGeneDrugPairs
3. Boxwala AA, et al. A multi-layered framework for disseminating knowledge for computer-based decision support. JAMIA 2011; 18:i132-i139.

Prioritizing Experimental Validations of Computational Predictions based on Estimated Biomedical Impact

Madhavi K. Ganapathiraju[1,2] Ph.D. Naoki Orii[1,2] M.S. and Lavanya Viswanathan B.Tch.[1,2]
[1]Department of Biomedical Informatics, University of Pittsburgh
[2]Language Technologies Institute, Carnegie Mellon University

Summary

We describe the need for a new line of algorithms in Translational Bioinformatics called "inference analytic" algorithms. These algorithms are expected to fit between bioinformatics computation and translation to biology or medicine, and are meant to carry out research prioritization so that the limited resources available for experimental or clinical work can be invested on those computational outcomes that are expected to have the biggest impact. We present application of this approach to the translational bioinformatics of protein-protein interactions.

Introduction and Background: There has been a steady increase in the number of biological hypotheses that are generated by translational bioinformatics (TBI) community. It is often infeasible to test all of the computationally generated hypotheses by biological or clinical methods as they would be resource-intensive. If only some of all the outcomes (hypotheses) of TBI algorithms are to be translated, how are those to be selected? One criterion can be to select those that are estimated to have biggest impact on future science. Note that 'impact on future science' itself can be defined in many ways – is it the biological, clinical or computational methods that have the impact – each choice would result in a different prioritization. Here, we present prioritization based on impact on biology, for a specific TBI algorithm of predicting protein-protein interactions (PPIs). PPIs are extensively used in molecular and systems biology, and our goal is to identify which of the computationally discovered PPIs will have the largest impact on biomedical science.

Methods: We presented the concept of impact prediction recently in [1]. We use number of citations as a surrogate measure for biomedical impact – although this may not be true in other domains, when we consider the papers that publish a single PPI and then look at the citations of that paper, this is a good indicator of impact of the PPI [1]. We present our results on impact prediction for PPI by adding Gene Ontology (GO) features, to the network-topology that we considered in our prior work. We included the 3rd-level GO terms separately for molecular function, biological process and cellular component. A random forest model is trained and tested on a dataset of all human PPIs, the publications reporting those PPIs and the citations received by those publications. We carried out a 10-fold cross validation on our training set using the random forest model and compared it against the random model, that assigned to interactions a score from a uniform distribution from [0,1].

S. No.	Symbol1	Symbol2	PMID reporting the PPI	Cita-tions
1	**HIST1H4A**	KDM4A	17190600	183
2	HIST1H3A	PRMT5	22231400	10
3	TCEB1	VHL	10449727	18
4	**PHYHIP**	PRMT5	16169070	451
5	RB1	**SPIB**	10196196	7
6	**NCOR2**	**HDAC4**	12205093	10
7	HIST1H3A	**PHF8**	20421419	11
8	**RPS6KB1**	EEF2K	11500364	104
9	NR3C1	SMARCE1	12917342	46
10	**CDK4**	**CDKN2C**	21988832	6
11	XPO1	SNUPN	22833565	1
12	**PRMT6**	HIST2H3A	18077460	35
13	MYD88	TLR2	15107846	53
14	KMT2A	**HIST1H4A**	20452361	12
15	PRMT1	**HIST1H4A**	22498736	0
16	PRMT1	**HIST4H4**	17264152	9
17	EZH2	HIST1H3A	16224021	73
18	EZH2	C7orf25	21900206	12
19	**DNMT3A**	HIST1H3A	19834512	46
20	**HDAC9**	**HDAC4**	12590135	19

Table 1: Protein interactions that are predicted to have high impact. GWAS genes are shown in bold.

Results and Discussion: Our model consistently outperforms the random method across random method. After thus evaluating the method, we trained a model with all available data (that was previously separated into training and test data), and predicted impact of all PPIs in the whole human interactome in order to identify high impact interactions. In the table are shown the top 20 interactions predicted as impactful by our model, that were not in our training set, along with their citation counts and score. It can be seen that some of them already achieved the high impact whereas there are some PPIs that have untapped potential for impact. This method would be useful when applied to computationally predicted PPIs or PPIs determined with high throughput technology that yield a large number of PPIs at once and taking them to wetlab needs prioritization Among the top predicted high impact interactions, there are those that are associated with diseases through GWAS (HDAC4, PHF8 etc), shown in bold in Table 1.

References

1. Ganapathiraju MK, Orii N: Research prioritization through prediction of future impact on biomedical science: a position paper on Inference-Analytics. GigaScience 2013, 2, 11.

The PhenX Toolkit: Promoting Data Sharing and Translational Research

Carol M. Hamilton, PhD[1]; Wayne Huggins, PhD, PBS[1]; Huaqin Pan, PhD, MS[1]; Elizabeth Eubanks, BA[1]; Deborah R. Maiese, MPA[1]; Destiney S. Nettles, MPM[1]; Joseph G. Pratt, MPM[1]; Tabitha P. Hendershot, BA[1]; Kevin P. Conway, PhD[2]; Kay L. Wanke, PhD, MPH[3]; Gregory K. Farber, PhD[4]; Erin M. Ramos, PhD, MPH[5]

1. RTI International, Research Triangle Park, NC
2. National Institute on Drug Abuse, National Institutes of Health (NIH), Bethesda, MD
3. Office of Disease Prevention, NIH
4. National Institute of Mental Health, NIH, Rockville, MD
5. National Human Genome Research Institute, NIH, Bethesda, MD

Abstract

*The PhenX (consensus measures for **Phen**otypes and e**X**posures) Toolkit (https://www.phenxtoolkit.org/) is a publicly available, online catalog of measures of phenotypes and exposures for use in genomic and epidemiologic research. The Toolkit has a broad scope, providing assessment protocols for 339 measures across 21 research domains, including, Demographics, Anthropometrics, Environmental Exposures, Cardiovascular, Diabetes, Neurology, Physical Activity and Physical Fitness, and Social Environments. The goal is to promote the use of standard measures, enhance data interoperability, and help investigators identify opportunities for collaborative and translational research. The content, functionality and maintenance of the PhenX Toolkit are driven by the scientific community. A steering committee (SC) provides overall guidance and establishes criteria for PhenX measures. Working groups (WG) of experts select measures for inclusion in the Toolkit. All SC and WG decisions are by consensus and consider input from the Institutes and Centers of the National Institutes of Health (NIH) and the broader scientific community. Approaches to integrate PhenX with established standards, data repositories and Electronic Medical Records (EMRs) will be presented. In collaboration with PFIND**R** (Phenotype Finder IN Data Resources) http://grants.nih.gov/grants/guide/rfa-files/RFA-HL-11-020.html PhenX measures are being mapped to all completed studies in the database of Genotypes and Phenotypes (dbGaP). Although the initial emphasis was to identify measures suitable for genome-wide association studies (GWAS), many PhenX measures are also suitable for use in a clinical environment and could be adapted for inclusion in EMRs. PhenX measures can be used to track the etiology and progression of numerous common complex diseases, such as diabetes, cardiovascular disease, obesity, and depression. Examples of PhenX interviewer-administered and self-administered questionnaires include demographics measures, family history of common diseases, and use of alcohol, tobacco, and other substances. The Toolkit includes 21 clinical exams, 33 physical measurements and 28 bioassays. Collections of measures for post-traumatic stress disorder (PTSD) and suicidality are currently in development; some of these measures may also be suitable for EMRs. Investigators and clinicians can find measures of interest by browsing or searching the Toolkit using the Smart Query Tool. For each measure, the Toolkit provides a description of the measure, the rationale for its inclusion, detailed protocol(s) for collecting the data, and supporting documentation. Toolkit users can download custom data collection worksheets and custom data dictionaries that support data submission to dbGaP. Currently, the dbGaP advanced search tool includes "PhenX" as a filter option and the dbGaP variable descriptions present dbGaP-PhenX variable classifications (identical, comparable or related). To further promote cross-study collaborations, the Toolkit includes a "Register Your Study" feature. Registered users can browse information about each registered study (principal investigator, research focus, number of subjects, study design) and see the PhenX measures that are included in each study; a summary of registered studies will be presented. We recognize that PhenX measures are complementary to other ongoing measures and standards initiatives. The presentation will also include an overview of available resources for recommended measures, common data elements and standards.*
Funding provided by U01 HG004597, 3U01 HG004597-02S1, and U41 HG007050.

COSMOS: NGS Analysis in the Cloud

Jared B. Hawkins[1], Ph.D., Yassine Souilmi[2], M.Sc., Ryan Powles[3], Jae-Yoon Jung[1], Ph.D., Alex K. Lancaster[1], Ph.D., Dennis Wall[4], Ph.D., Peter Tonellato[1], Ph.D.
[1]Harvard Medical School, Boston, MA, [2]Faculty of Sciences of Rabat, Morocco, [3]Virginia Tech, Blacksburgh, VA, [4]Stanford University, Stanford, CA

Summary

COSMOS is a powerful workflow management system that enables the timely and cost-effective implementation of complex next-generation sequencing (NGS) analysis pipelines. COSMOS takes advantage of parallelization and the resources of a high-performance compute cluster, either local or in the cloud, to process datasets of up to the petabyte scale, which is becoming standard in NGS.

Introduction

NGS costs have plummeted in recent years, rapidly outpacing the traditional benchmark for the decreasing cost of technology known as Moore's law. Modern sequencing platforms are capable of sequencing approximately 5,000 megabases a day. As a result, programs such as the 1000 Genomes project (Broad Institute) are routinely generating data on the petabyte scale. The current challenge lies in the analysis and interpretation of this data, which has become the new rate-limiting step. Providing a solution to this problem that is both timely and cost-effective will be of great scientific importance and a major technological breakthrough. To address this need, we have developed a scalable, parallelizable workflow manager capable of running on the cloud (e.g., Amazon Web Services – AWS) that has the potential to reduce the cost of the analysis of whole genomic data over 10-fold, from ~$1,000 to under $100. COSMOS is able to lower the cost of analyzing genomic data two ways: 1) implementing a highly parallelizable workflow that can be run quickly and efficiently on a large compute cluster, and 2) taking advantage of AWS spot-instance pricing to reduce the cost per hour.

Methods

COSMOS is a workflow manager for which we have developed a NGS-analysis pipeline called GenomeKey. Genomekey implements the Genome Analysis Toolkit (GATK) best practice protocol developed by the Broad Institute, which is widely accepted as the industry standard. Our pipeline performs a thorough sequence analysis, including quality control, alignment, variant calling and an unprecedented level of annotation using a custom extension of ANNOVAR. After loading in genomic data, COSMOS breaks up each stage of the workflow into multiple jobs that can be efficiently run in parallel. Jobs are distributed to worker nodes within the cluster using a standard job manager (e.g. we use Sun Grid Engine on AWS).

Results and Discussion

By using the COSMOS framework to process and analyze successively larger sets of genomic data, we have optimized the software for speed, accuracy, simplicity, automation, and computational cost. In this way, COSMOS is a scalable, efficient solution to the increasing demand for effective computational systems in bioinformatics. Figure 1 demonstrates how COSMOS breaks up stages of the NGS-pipeline into multiple jobs for different sized datasets (shown here for 1, 5 and 10 genomes). This technique allows for efficient parallelization of the early stages of the pipeline (i.e., alignment). Later stages (i.e. variant calling) are run in serial to maximize accuracy of the analysis. To demonstrate the scalability of COSMOS, we are in the process of analyzing three separate cohorts: 1) 6000 exomes from Autism patients, 2) 100 genomes from Autism patients, and 3) ~300 genomes from ancient human DNA samples. We will present data from these analyses at the AMIA TBI 2014. Our presentation will focus on benchmarking and AWS-specific optimization to illustrate that COSMOS is a rapid, scalable and cost-effective tool for genomic data analysis in the cloud.

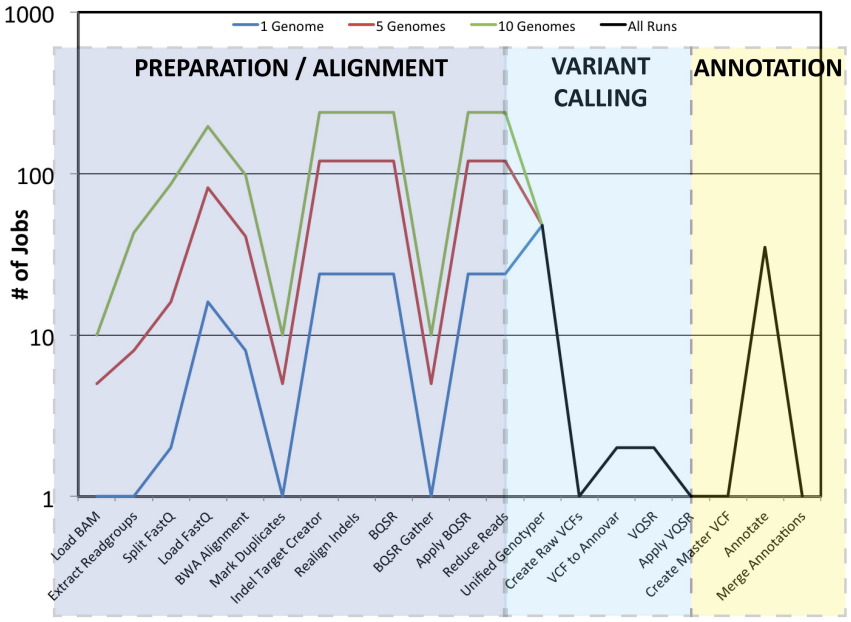


Figure 1. Parallelization of jobs for scalable datasets.

Improving Translatability of Biological Networks for Applications in Human Disease and Pharmacology

Alexandra Jacunski, MS[1]; Scott J. Dixon, PhD[2]; Brent R Stockwell, PhD[2]; Nicholas P Tatonetti, PhD[3,4,5]
[1]Integrated Program in Cellular, Molecular, and Biomedical Studies; [2]Department of Biological Sciences; [3-5]Departments of Biomedical Informatics, Systems Biology, and Medicine, Columbia University, New York, NY

Summary: A key challenge in biomedical research is translating relationships from model systems to humans. Even well-studied properties, such as genetic interactions, are poor predictors of human counterparts. We hypothesize that biological network properties can circumvent this problem. We develop a generalized algorithm for interspecies network translation and apply it to predict and explore synthetic lethal (SL) interactions in human cells.

Introduction: Network medicine employs biological networks to contextualize systems through computational methods, and has been used to investigate complex and infectious diseases, such as cancer, and to predict novel yeast SL interactions. SL is a genetic interaction wherein simultaneously knocking out two nonessential genes causes inviability of a cell or organism. While heavily investigated in *S. cerevisiae*, SL has only recently come of interest in human cancer research. Because most drugs abrogate gene function, it stands to reason that drugs affecting two SL genes can be synergistically exploited in chemotherapy.

We hypothesized that translating information from biological networks will allow us to use data from *S. cerevisiae* to predict SL in humans, where it is understudied. We explored these predictions using drug-drug interactions (DDIs) to identify potential novel synergies. We validated our findings experimentally with small-molecule inhibitors in three predicted SL pairs.

Methods: We constructed protein-protein interaction (PPI) networks for *S. cerevisiae, S. pombe, H. sapiens,* and *M. musculus*, then calculated and translated graph parameters for each network. We used these data and SL status from *S. cerevisiae*, our source species, to train a random forest classifier. The classifier was applied to the data of a target species, first validating in *S. pombe*.

Predicted human SL pairs (Pr(SL)>95%) were then mapped to drugs using the Drug Combination Database (DCDB) and DrugBank. We experimentally investigated previously undocumented synergistic DDIs affecting *BRAF* and the *MAP2K* genes by measuring lethality in HT1080 cells from small molecule inhibitor combinations.

We then began exploring environmental changes in the PPI network by integrating gene expression data. We hypothesized that such changes may cause gains or losses of SL relationships between genes, using normal and peroxide-stressed *S. cerevisiae* networks. We expect that this method can help predict differential adverse drug reactions (ADR) in tissue- or disease-specific human networks.

Results: Our classifier successfully predicted SL in *S. cerevisiae* gene pairs (AUC = 0.92, raw and translated data). We validated its success in translation using *S. pombe*, attaining AUCs of 0.63 and 0.86 using raw versus rank-normalized data (Figure 1). We then applied it to the *H. sapiens* network. Mapping our human SL predictions (Pr(SL) > 0.90) to drugs in DCDB yielded known DDIs for 61% of gene pairs. Additionally, SL pairs (Pr(SL) > 95%) mapped to DrugBank produced three gene pairs, all associated with at least one DDI.

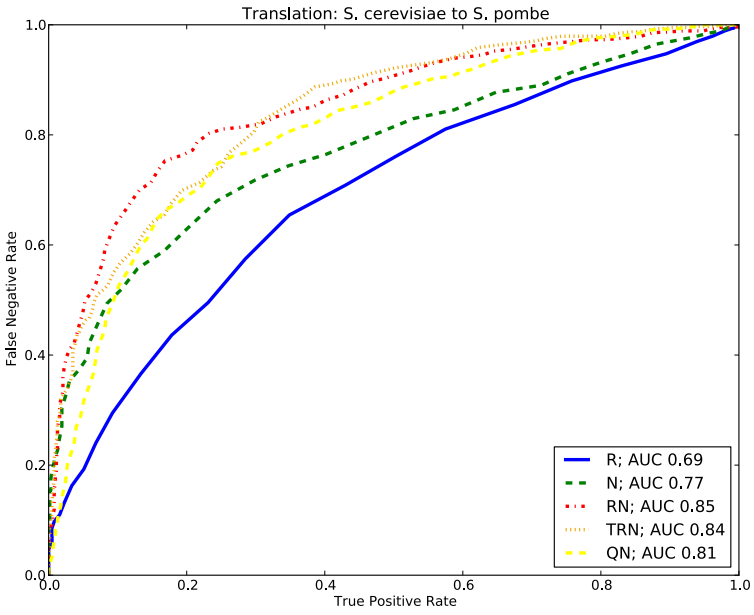


Figure 1: Prediction of SL from *S. cerevisiae* to *S. pombe*. Translation methods: R = raw; N = normalized. RN = rank-normalized; TRN = tied-rank normalized; QN = quantile normalized.

We observed that only three DCDB pairs with Pr(SL)>95% mapped to drug targets: *BRAF/MAP2K1, BRAF/MAP2K2,* and *MAP2K1/MAP2K2*. We assayed these interactions by measuring cell death after exposure to their small molecule inhibitors in combination. The observed synergies strongly support our assertion that these pairs are SL.

We then constructed PPI networks with edges weighted by gene coexpression in normal and stressed *S. cerevisiae*. Since studies have found that environmental changes alter PPI networks, we hypothesize that SL relationships may also change.

Discussion: Our results indicate successful translation between biological networks, allowing for improved information exchange between model organisms and humans. In particular, our applications of this methodology have elucidated SL gene pairs in humans. SL predictions from general and environmentally-altered PPI networks can be used to anticipate DDIs.

Identification of Transcriptionally-Defined Cancer Subpopulations Through Integration of Public Microarray Data with Single Cell Gene Expression Profiling

Michael Januszyk, MD[1], Michael Sorkin, MD[1], Robert C. Rennert, MD[1], Geoffrey C. Gurtner, MD[1], Purvesh Khatri, MD[1], Atul J. Butte, MD, PhD[1]
[1]Stanford University, Stanford, CA

Abstract
Complex cell populations are characterized by a level of transcriptional heterogeneity that likely underlies their biological function. Here we present a method to analyze such complex cell types combining the knowledge derived from public microarray datasets with the precision afforded by single cell analysis, applied to study lung cancer.

Background
Over the last fifteen years DNA microarray analysis has yielded tremendous insight into the nature of cancer cell populations, leading to numerous advances in treatment strategies. However, it has become increasingly clear that these tumors are highly heterogeneous, with subsets of cancer stem cells believed to be responsible for the tumorigenic capacity of each parent population. Single cell gene expression analysis represents an attractive approach to investigate such complex cell populations for which the granularity afforded by traditional microarray-based analyses is generally insufficient. However, these high-resolution techniques are typically limited in the number of gene targets that may be simultaneously interrogated (generally less than 100) and therefore capture only a small fraction of each cell's transcriptome. This limitation may be addressed by screening public microarray data in order to generate an informed list of genes likely to be differentially expressed for each cancer subtype.

Methods
A meta-analysis was performed across six public microarray datasets for human small cell lung cancer (SCLC) comprising 365 samples across eight different platforms. Genes were ranked according to effect size and p-value for tumor versus control samples, and false discovery rates were calculated. The top scoring 48 genes that were significant by both methods, along with the 48 highest rated surface antigen genes, were used to populate a gene list for subsequent single cell evaluation. High throughput gene expression analysis was performed for 400 individual cells from one SCLC line (H446) using the Fluidigm microfluidic platform.

Results
Supervised machine learning was applied to identify transcriptionally-defined subgroups among SCLC cells, and hierarchical clustering performed against cells within each cluster as well as across all 96 genes. Three distinct subgroups were identified using K-means clustering, and one cluster exhibited increased expression of genes associated with aggressive malignancy such as PDIA, JAG1, and CD47. The non-parametric Kolmogorov-Smirnov test was then used to compare the expression of each surface marker gene between cells in this cluster of interest and the remaining cells. FACS analysis was then performed for the top scoring markers (DDR1 [$p = 2.33e-09$] and CD47 [$p = 3.38e-08$]) on SCLC cells. Two distinct populations were observed when sorting for DDR1. Three sub-populations of SCLC cells (DDR1+, DDR1-, and unsorted) were prospectively isolated using FACS and evaluated in vitro in culture. Single cells were sorted into 48 well plates and clones were evaluated after 14 days in culture. DDR1+ cells demonstrated significantly greater clonogenic capacity than either DDR1- or unsorted cells, suggesting that this marker may enrich for a more aggressive subset of SCLC cells.

Discussion
Meta-analysis of publicly available microarray data represents a unique opportunity to analyze heterogeneous datasets to identify subtle yet significant changes in gene expression across all cohorts. Such an analysis can complement single cell profiling technologies by generating informed gene lists that can be used to detect differences within heterogeneous populations. Identification of novel subtypes among complex cell populations may facilitate the development of new diagnostic and therapeutic techniques centered on the enrichment or depletion of specific subpopulations with distinct functional properties.

INTRA-HOST EVOLUTION AND SUBCLONAL DIVERSITY IN ACUTE INFECTIONS USING UNPHASED GENOMIC DATA

Hossein Khiabanian, PhD [1, 2], Zachary Carpenter, MS [1], Jeffrey Kugelman, PhD [3], Joseph Chan, PhD [1], Vladimir Trifonov, PhD [1], Elyse Nagle, PhD [3], Travis Warren, PhD [3], Patrick Iversen, PhD [4], Sina Bavari, PhD [3], Gustavo Palacios, PhD [3], Raul Rabadan, PhD [1, 2]

[1] Department of Systems Biology, [2] Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons, New York, NY 10032; [3] Genomics Division, The U.S. Army Medical Research Institute of Infectious Diseases, Fort Detrick, MD 21702; [4] Discovery Unit, Sarepta Therapeutics, Corvallis, OR 97333

Summary. A single highly evolving virus can give rise to a swarm of related descendants. Utilizing deep-sequencing data, we introduce two measures of diversity to infer times to the most recent common ancestor, evolutionary rates, and the presence of selection during acute infections. Our method is able to identify subclonal events present in <1% of the population, capturing genomic diversification within days of an infection.

Introduction. Intra-host viral diversity has been postulated to enable a virus to explore larger sections of the fitness landscape, driving immune evasion and drug resistance. Genomic approaches based on consensus sequences that have been used to study inter-host pathogen transmission, fail to capture subclonal diversity of intra-host viral populations. However, high-throughput sequencing technologies provide us a glimpse into this diversity, despite the inability to assess whether two short reads originate from the same individual clone, a problem known as phasing. **Methods.** Utilizing unphased genomic data, we introduce measures of diversity based on i) total divergence, D_T, the sum of frequencies of diverging alleles from the original clone, and ii) the sum of minimal allele frequencies (MAF) at segregating sites. Strong differences between the two measures indicate selection or bottlenecks, as changes in D_T measure time and divergence from the seed and the sum of MAF indicates variations in population diversity at a particular time. Assuming a molecular clock, the number of mutations acquired in each individual during a clonal expansion can be approximated as Poisson distributions with a mean proportional to the length of the genome, evolutionary rate, and the time that has passed since the start of the expansion. Therefore, given sets of sequences collected at various time points, the evolutionary rate and the starting time of an expansion can be estimated by maximizing the product probability of respective Poisson distributions. We follow a maximum likelihood estimation (MLE) approach, and assume negligible back mutations and initial infection by a single viral clone. This methodology holds regardless of recombination or reassortment events within a clonal population because such evolutionary processes do not affect genetic diversity. This method is specifically applicable to temporal intra-host genomic data obtained from acute infection by lytic viruses; however, it cannot be applied to integrating or lysogenic viruses, such as HIV in its chronic phase. In addition, seasonality can be inferred from spectral analysis of the diversity measures. **Results.** We evaluate our methodology with well-characterized influenza pandemics and epidemics, for which our findings corroborate previous analyses. More importantly, we apply our method to temporal data obtained by ultra-deep sequencing of intra-host viral populations from four marburgvirus-infected non-human primates (cynomolgus macaque). These independent analyses show i) an increasing genomic diversity with a rate, μ, of 4.8 x 10^{-4} and 10.1-26.5 x 10^{-4} substitutions/site/year for non-synonymous and all substitutions, respectively, and ii) 3-11 days to convergence with the reference, t_0, approximately the amount of time spent propagating the virus after the original sequencing of the seed of the infection. The ratio of non-synonymous to synonymous substitutions rates was found bellow 1, indicating strong purifying selection, but higher rate of evolution at intra-host levels compared to inter-host transmission of marburgvirus. **Conclusions.** Our method is able to analyze regional and subclonal genomic diversification within days of an infection. Thus, it is an ideal tool of benchmarking, particularly in therapeutic studies of expanding populations, to pinpoint the time of infection, to estimate the evolutionary rate within a host, and to study acute RNA virus infection dynamics.

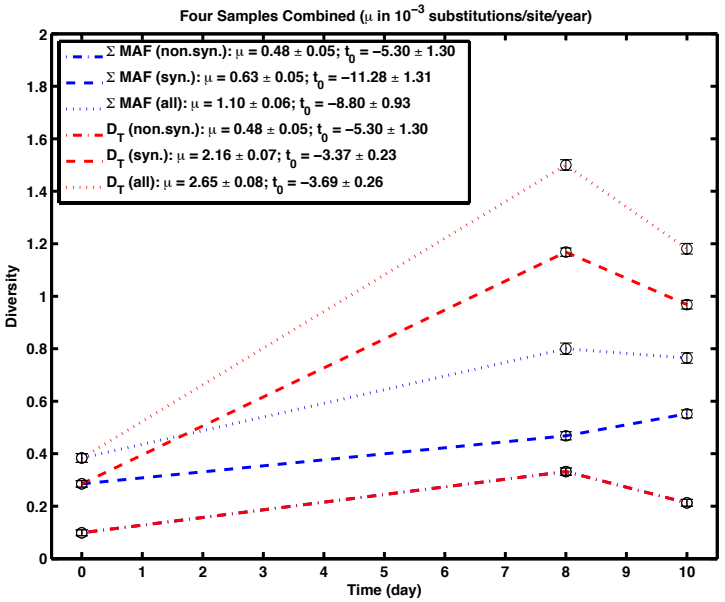


Figure: MLE results for combined data from all four marburgvirus samples. The standard errors are derived from 95% confidence intervals via bootstrapping

Towards an Integrated Framework of Pharmacovigilance Signal Detectors through Semantic Mediation

Vassilis G. Koutkias, PhD[1] and Marie-Christine Jaulent, PhD[1]
[1]INSERM, U1142, LIMICS, F-75006, Paris, France; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France

Abstract

We present our initial steps for developing an integrated framework of pharmacovigilance signal detection methods exploiting diverse data sources. Our approach lies on semantic mediation among heterogeneous signal detectors and, at this stage, we focus on constructing an ontology for their semantic description. Our ultimate goal is to complement evidence on potential signals to support domain experts in more accurate, timely and reliable detection.

Introduction & Background

A major challenge in the field of pharmacovigilance is the accurate identification of drug safety signals, i.e. information on possible causal relationships between drugs and adverse effects. Various statistical and knowledge inference techniques have been proposed for signal detection, according to the type of data sources being explored. However, current detection methods demonstrate high-rate of false-positive indications, as well as variability in their results depending on the analysis parameters set. Considering the difficulties in detecting and assessing potential drug safety risks, the combined exploitation of all the evidence is required in order to leverage signal detection[1,2].

Methods

Our research focuses on the construction of an integrated framework for pharmacovigilance signal detection. The detectors considered in this work exploit data from Spontaneous Reporting Systems (SRS), Electronic Health Records and text-based sources (e.g. scientific papers and social media platforms). Based on the literature and hands-on experience, we analyzed the functional characteristics as well as the strengths and limitations of currently available signal detectors. For example, the performance of SRS-based detectors is affected by inconsistent reporting, latency and reporting bias, while detectors exploring social media lack quality control and exploit unstructured data. The realization of the proposed framework lies on semantic mediation among detectors through an ontology that we are developing for describing their characteristics in terms of input, analysis options and output.

Results

The design of our ontology has been influenced by the IOPE (Inputs-Outputs-Preconditions-Effects) model, employed in OWL-S (http://www.w3.org/Submission/OWL-S/). It also relies on concepts linked with drug safety risks. At this stage, the major focus of our implementation has been given on detectors that are available via OMOP (Observational Medical Outcomes Partnership, http://omop.org/). Having as the basis semantic technologies and service-oriented computing, the technical requirements of the framework are being specified, aspiring to define an open platform for concurrently gathering and synthesizing the outcome of diverse signal detection methods.

Discussion

The identification of pharmacovigilance signals is laborious and time-consuming, since the indications provided by current signal detection tools are highly uncertain and need to be further assessed. Our next steps will rely on a fully functional integration using our ontology, concentrating on ways to automatically align, compare and potentially complement evidence provided by diverse, available signal detectors (either in-house, or open-source software).

Acknowledgement

This research was supported by a Marie Curie Intra European Fellowship within the 7th European Community Framework Programme FP7/2007-2013 under REA grant agreement n° 330422 – the SAFER project.

References

1. Kelman CW, et al. Evaluating medicines: let's use all the evidence. Med J Aust 2007;186(5):249-52.
2. Hauben M, Norén GN. A decade of data mining and still counting. Drug Saf 2010;33(7):527-34.

Validation and Portability of Unbiased, Label-free Proteomics

J. Will Thompson[2] Ph. D., Laura Dubois[2], Keyur Patel[2] M. D., M. Arthur Moseley[2] Ph. D., Joseph E. Lucas, Ph. D.[1,2]

[1]Quintiles, Morrisville, NC; [2]Duke University, Durham, NC, USA

Studies utilizing unbiased, label-free "shotgun" proteomics are often difficult to validate. Inability to identify some peptides in the sample and challenges with matching detected peaks across samples and experiments can lead to failed attempts to validate proteomics findings. The problem is particularly difficult when the number of peptides targeted for validation is small. We have developed a factor model approach to the analysis of shotgun proteomics data; This approach generates results that can be validated with further shotgun proteomics experiments and leads to a novel technique for the identification of targets for validation by selected reaction monitoring. We demonstrate the use of our analytic approach in four different Hepatitis C data sets, including (1) discovery (n=55) in a time of flight mass spectrometer, (2) validation (n=41), (3) validation in a different lab with an orbitrap mass spectrometer (n=35), and (4) selection of targets and validation in a much larger cohort of patients (n=149) using selected reaction monitoring (SRM).

Methods. We utilized a Bayesian factor model tailored to take advantage of known features of proteomics data: identification of only a subset of measured peptides, high levels of correlation between peptides in the same protein, and high levels of correlation between proteins in the same biological pathway. The resulting factors, which are each representations of the expression of multiple peptides, were then used as independent variables to predict response to therapy in patients with Hepatitis C. Because the factor model results in significant dimension reduction, we were able to build models that were dependent on multiple predictors.

We match peaks between experiments with a greedy algorithm based on agreement between retention time and M/Z in the two data sets. This approach leads to probable matches with some mistakes. However, because each factor in our predictor is approximated by many measured peptides, failure of some of these peptides to match across data sets or to validate leads to only minor changes in the accuracy of the predictor. This allows us to validate the predictor in both a new set of samples and in data generated by a different lab with different technology.

Selection of targets for SRM is typically conducted based on the assumption that every peptide from a particular protein is reporting on the expression changes of that protein. However, we show that this is an oversimplification; Proteins quite often have peptides with expression patterns that are apparently uncorrelated – or even negatively correlated. Based on the factor model and factor regression, we derive an expression that describes the loss in accuracy of the predictor that occurs because a subset of peptides is not measured. Selection of SRM targets is then a matter of minimizing the loss in accuracy under the constraint that we measure a limited number of peptides.

Results. Training of both factor model and factor regression predictor was conducted on pre-treatment serum samples from a cohort of 55 Hepatitis C patients with known response to standard of care (Ribavirin and Interferon). Area under the receiver operating characteristic (AUROC) for this training data was 0.9. AUROC on a validation cohort of 41 pediatric patients with Hepatitis C was .8 (p-value 4e-4) and AUROC on 35 samples which were measured in a different lab was 0.84. Subsequently, the training and pediatric data sets were used to identify peptides for SRM. Additional samples from these cohorts were then measured by SRM to train a new model based on just 10 peptides. This model was validated in a new cohort of 149 hepatitis C patients resulting in an AUROC of .73 (p-value 1.6e-5).

Discussion. By allowing high throughput measurement of thousands of peptides simultaneously, unbiased, label-free proteomics studies can provide a window into disease processes that is inaccessible by any other technique. However, challenges with data analysis have made translation of discoveries into the clinic difficult. We have developed a statistical platform that can help to bridge the gap between discovery and clinical implementation, and have demonstrated its use from discovery through clinical validation. This approach offers a smooth path for future unbiased proteomics studies to move into the clinic and thereby result in a real impact on patient care.

Efficient Algebraic Interval Queries on Biomedical Sequence Annotations

Yuan Luo MS, Peter Szolovits, PhD

Massachusetts Institute of Technology, Cambridge, MA

Abstract

We present an algorithmic framework based on augmented interval tree for solving algebraic interval queries on biomedical sequence annotations with optimal time complexity.

High throughput technologies yield vast volume of data that often comes at high velocity in different biomedical subdomains including genomic sequencing analysis and clinical time series analysis. In addition, many clinical natural language processing (NLP) systems employ stand-off annotations aligned by text coordinates that are used to index large numbers of electronic medical records (EMRs). Although individual problems differ in the nature of their data, these problems share common structure in that all can be abstracted to the interval storage and query problem. Thus, the ability to efficiently store, update and query the intervals is under increasingly pressing demand.

Previous works on applying interval indexing in the biomedical field limit their attention to the overlapping queries. On the other hand, to better characterize the relations between different intervals, Allen [1] proposed the now widely accepted theory of interval algebra, which includes 13 interval relations listed below, where i, j are intervals. Although originally proposed for calculating temporal logic, Allen's interval algebra generalizes naturally to other sequence annotations. In this work, we propose an efficient interval tree based algorithmic framework for querying intervals using each of the 13 interval relation given a query interval.

$i = j$	i is **equal** to j
$i < j$ (>)[1]	i is completely to the **left** of j
$i\ m\ j$ (mi)	i's ending point **meets** j's beginning point
$i\ d\ j$ (di)	i is **during** j
$i\ s\ j$ (si)	i **starts** same as j and finishes before j
$i\ f\ j$ (fi)	i **finishes** same as j and starts after j
$i\ o\ j$ (oi)	i **overlaps** j and starts before j

We call the problem of retrieving all intervals satisfying certain relation with a given interval (point) as a stabbing interval (point) query. Finding the interval with max weight containing the query point is called a stabbing-max point query. The state-of-the-art interval tree implementation in the biomedical domain [2] relies on the basic interval tree. Its stabbing interval query finds intervals that share common region with a given interval, which translates to "$o \vee oi \vee s \vee si \vee d \vee di \vee f \vee fi \vee =$" in Allen's algebra. In practice, stabbing interval queries on fine grained relations are often desirable. For example, most existing genetic databases assign their own customized identifiers to genetic sequences at various levels including locus, transcript, and probe. To search for mutations within the breast cancer early onset gene *brac2*, one needs to rely on genomic intervals and perform a "d" query given the interval of *brac2*. However, we prove that the query time used by the basic interval tree on relations such as "d", "o" and "oi" has a worst case complexity of $O(log^2 n)$, far from being optimal. We address this problem by proposing an augmented interval tree with optimal stabbing max-point query time, insertion time, and updating time at the same time. The key idea is to embed a secondary tournament tree for each node in the basic interval tree to keep the intervals associated with that node in sorted fashion so that retrieving the max is efficient without sacrificing insertion and updating time complexity. We then reformulate queries on difficult relations in Allen's interval algebra as stabbing max-point queries, as shown in Table 1.

Table 1. Reformulations for stabbing interval queries on difficult relations in Allen's interval algebra. The interval s is the reference interval. $w(.)$ indicates interval weight.

Allen's relation	Reformulation
$s = [x, y]\ o\ s' = [x', y']$	$y \in s'$ and $w(s') = x' > x$
$s = [x, y]\ oi\ s' = [x', y']$	$x \in s'$ and $w(s') = y' > x$
$s = [x, y]\ d\ s' = [x', y']$	$y \in s'$ and $w(s') = -x' > -x$

Complexity analysis shows that our interval tree framework attains the optimal stabbing interval query time complexity $O(log\ n + k)$ on all relations in Allen's algebra, as well as the optimal time complexity $O(log\ n)$ for insertion and updating on the tree, where n is the number of tree nodes and k is the size of the output.

References

[1] Allen, J.F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*. 26, (1983), 832–843.

[2] Mohammad, F. et al. 2012. AbsIDconvert: An absolute approach for converting genetic identifiers at different granularities. *BMC bioinformatics*. 13, (2012), 229.

[1] Relation in parentheses denotes the inverse relation.

ARTS: Automated Randomization of Multiple Traits for Study Design

Mark Maienschein-Cline[1,2], Zhengdeng Lei[1,2], Vincent Gardeux[2,3], Neil Bahroos[1,2], and Yves Lussier[2,3,4]

[1]Center for Clinical and Translational Science, University of Illinois at Chicago
[2]Institute for Interventional Health Informatics, University of Illinois at Chicago, Chicago, IL
[3]Departments of Medicine and Bioengineering, University of Illinois at Chicago, Chicago, IL
[4]Computational Institute of The University of Chicago and Argonne National Laboratory, Chicago and Lemont, IL

Abstract

Collecting data from large studies on high-throughput platforms, such as microarray or next-generation sequencing, typically requires processing samples in batches. There are often systematic but unpredictable biases from batch-to-batch, so proper randomization of biologically relevant traits across batches is crucial for distinguishing true biological differences from experimental artifacts. When a large number of traits are biologically relevant, as is common for clinical studies of patients with varying sex, age, genotype, and medical background, proper randomization can be extremely difficult to prepare by hand, especially because traits may affect biological inferences, such as differential expression, in a combinatorial manner. We will present ARTS (Automated Randomization of multiple Traits for Study design), which automatically optimizes batch assignment for any number of samples, any number of traits, and any batch size. Researchers may access ARTS via a downloadable command-line tool, as well as at the Galaxy installation hosted by the UIC Center for Research Informatics (CRI) at galaxy.cri.uic.edu.

Introduction

Data collected on high-throughput biological platforms, such as microarray and next-generation sequencing (NGS), can often be processed in parallel in batches, greatly lowering the cost and time for collection. However, details in the personnel, protocol, or instrument setting/calibration often vary slightly from batch-to-batch. When large studies with hundreds or thousands of samples are conducted, these variations may result in statistically significant, but biologically irrelevant, anomalies between batches[1], confounding efforts to determine true biological differences between sample conditions. Such batch effects can be mitigated by proper randomization[2], where sample traits, such as diseased or control, are evenly distributed across batches.

Here, we present the ARTS (Automated Randomization of multiple Traits for Study design) computational tool for automated study randomization, which can be applied to a study of any size, with any number of traits and any batch size. ARTS uses a genetic algorithm to optimize an objective function based on a rigorous statistic from information theory, mutual information. To validate ARTS, we test several objective functions to illustrate the versatility of the one chosen, and we show that the genetic algorithm we use for optimization obtains a good balance between computational speed and optimization quality.

Methods and Results

ARTS essentially consists of two parts: the objective function used to quantify randomization, and the algorithm used to optimize that objective function. We will first provide a theoretical and practical motivation for our choice of objective function. Then we will describe the genetic algorithm we use to optimize this objective function, and compare its performance to a brute force enumeration approach and a simpler Monte Carlo sampling procedure.

Second, we will then discuss a real-life use case that motivated the development of ARTS, where we constructed a randomized study design for samples from ~200 patients over 8 traits and batch sizes varying from 4 to 96 samples. We will also describe the availability of the method to researchers.

This work is supported in part by NIH grant UL1TR000050 (University of Illinois CTSA).

References

1. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Rev. Genet.*, 2010; 11, 733-739.
2. Hu, J., Coombes, K.R., Morris, J.S., and Baggerly, K.A. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief. Funct. Genomic. Proteomic.,* 2005; 3, 322-331.

Focused Proteomic Profiling for Late Onset Neonatal Sepsis

Subramani Mani MBBS PhD[1], Daniel Cannon MS[1], Carol Hartenberger RN[1],
Karri Ballard PhD[2], Hannah Peceny BS[1], Robin Ohls MD[1]
[1]University of New Mexico Health Sciences Center, Albuquerque, NM, 87131
[2]Myriad RBM, Austin, TX 78759

Discovering biomarkers for neonatal sepsis is a definitive clinical need. Our objective was identification of a set of protein biomarkers for rapidly and reliably detecting late onset sepsis. We enrolled 139 infants, performed a focused proteomic assay of 90 potential biomarkers and identified a set of five biomarkers with high predictive performance.

Introduction and background: Neonatal sepsis is a serious infection of newborns leading quickly to organ failure and death without prompt detection and treatment. The incidence of sepsis among infants weighing <1,500 grams is approximately 20%, 200-fold higher compared to term infants.[1] Late onset sepsis (LOS) occurring in infants who are ≥5 days of age is more than ten times common in infants admitted to the neonatal intensive care unit (NICU) when compared with early onset sepsis and is usually due to healthcare interventions. Because of its frequency and high risk of morbidity and mortality, 'rule out sepsis' accounts for more than half of admission diagnoses made in the NICU.[2] A specific biomarker or a set of biomarkers that could clearly identify infection in low birth weight and premature infants and thereby aid in the early detection and management of LOS is a definitive clinical need. Our objective was identification of a set of ranked protein biomarkers for rapidly and reliably detecting LOS.

Methods: We enrolled 139 infants over a five year period (2007-2012) based on the following inclusion criteria: 1. gestational age of ≤32 weeks; 2. birth weight ≤1500 grams and 3. postnatal age of ≥5 days. Out of 139 infants 46 were cases and 93 were controls. Only culture positive infants were included as cases in our study sample. Apart from demographic, clinical and routine lab measurements, a focused proteomic assay of 90 potential protein biomarkers suspected to play a role in infection and inflammation was also performed using serum samples drawn over a 21-day period. For cases we define t_0 as the day blood was drawn that resulted in the infant's first positive culture and for controls t_0 was selected to match the day of age of infected infants. For prediction we used only data that was collected using samples drawn on and before t_0. We used decision tree based machine learning algorithms such as random forests (RF), C4.5, CART and reduced error pruning (REP) and a correlation based feature (variable) selection algorithm CFS to obtain a set of predictive protein markers and rank them.

Table 1. Predictive performance of ML models for sepsis using 10 x 10-fold cross validation and feature selection.

Algo	AUC (s.d)	Accu. (s.d)	Sens. (s.d)	Spec. (s.d)	PPV (s.d)	NPV (s.d)
C4.5	0.73 (0.03)	0.80 (0.02)	0.61 (0.04)	0.90 (0.03)	0.74 (0.05)	0.82 (0.02)
CART	0.75 (0.02)	0.83 (0.01)	0.66 (0.02)	0.91 (0.02)	0.79 (0.03)	0.84 (0.01)
RF	0.85 (0.01)	0.83 (0.01)	0.65 (0.02)	0.92 (0.02)	0.81 (0.03)	0.84 (0.01)
REP	0.73 (0.04)	0.80 (0.02)	0.59 (0.03)	0.91 (0.02)	0.76 (0.05)	0.82 (0.01)

Results: RF had the best overall predictive performance with an AUC of 0.85, sensitivity of 0.65, specificity of 0.92, positive predictive value (PPV) of 0.81 and a negative predictive value (NPV) of 0.84 (see Table 1). The top five predictive biomarkers based on RF variable importance ranking were 1. Tumor necrosis factor receptor 2 (TNF R2), 2. Interleukin-6 (IL-6), 3. Calcitonin, 4. C-Reactive Protein (CRP) and 5. Granulocyte Colony Stimulating Factor (G-CSF).

Discussion and Conclusion: Five biomarkers were identified that predicted infection with high accuracy. Additional validation of the models is necessary before clinical use is possible. We plan to combine clinical data with proteomic data to improve performance. The specific biomarkers TNF R2, IL-6, Calcitonin and CRP are known to be associated with inflammation and infection. However, only Calcitonin and CRP are used clinically, and only CRP is routinely used in neonatal clinical practice. This study opens up the possibility of a proteomics-based diagnostic approach for early detection of late onset neonatal sepsis in preterm infants.

References

1. Stoll BJ, Hansen N, Fanaroff AA, Wright LL, Carlo WA, Ehrenkranz RA, et al. Changes in Pathogens Causing Early-Onset Sepsis in Very-Low-Birth-Weight Infants. *New England Journal of Medicine* 2002b;347(4):240-47.
2. Gerdes JS, Polin RA. Sepsis screen in neonates with evaluation of plasma fibronectin. *The Pediatric Infectious Disease Journal* 1987;6(5):443-46.

Informatics for the International Mouse Phenotyping Consortium

Terrence F. Meehan[1] on behalf of the MPI2 Consortium

1= European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

The **International Mouse Phenotyping Consortium** (IMPC) is building the first truly comprehensive functional catalogue of a mammalian genome. This global effort requires generating and characterizing a knockout mutant strain for every protein-coding gene in the mouse. Data from a standardized, broad-based phenotyping pipeline are collected and archived centrally by the IMPC-Data Coordinating Center, including legacy data from pilot projects, EUMODIC and Sanger Mouse Genetics Project. Dedicated 'data wranglers' are working with each phenotyping center to ensure proper transfer and quality control of data. An automated statistical analysis pipeline identifies knockout strains with significant changes in phenotype parameters. Annotation with biomedical ontologies allows biologists and clinicians to easily find mouse strains with phenotypic traits relevant to their research. Data integration with other resources provides insights into mammalian gene function and human disease. Users can freely access all data including new gene-phenotype associations via APIs and an intuitive web portal. The community is invited to explore and provide feedback as we build this rich resource for precision medicine at:
www.mousephenotype.org

Physiological Predictors based on Temporal Clustering of Patients

Robert Moskovitch, PhD[1] and Nicholas P. Tatonetti, PhD[1]
[1]Departments of Biomedical Informatics, Systems Biology, and Medicine
Columbia University, New York, NY;

A patient's health status is a dynamic process of diagnosed diseases, prescribed treatments, and managed reactions. The emergence of temporal clinical data (e.g. EHR) enables novel predictive models of these processes. Here we present a hybrid approach that integrates temporal pattern discovery with a-temporal predictive features.

Introduction and Background

The increasing availability and accumulation of temporal clinical and patient data provide exceptional opportunities for translational biomedicine, including temporal-based pattern discovery, patient clustering, and predictive modeling. A major challenge in biomedical informatics is correlating patient outcomes to their physiological and molecular properties [1]. Most methods in time series analysis focus on univariate time series, while others expect the data to be sampled uniformly. However, biomedical data along time contain heterogeneous data types, are non-randomly missing, and inconsistently sampled [2]. To address these challenges Temporal Abstraction (TA) [3], which is the segmentation and/or aggregation of a series of raw, time-stamped, multivariate data into symbolic time-intervals series representation, was proposed as shown in figure 1. These models enable prediction of target events based on the evolution of the data along time. However, an even more powerful approach is to leverage temporal features to identify a-temporal predictors, such as age, gender and other static properties, such as generic variants.

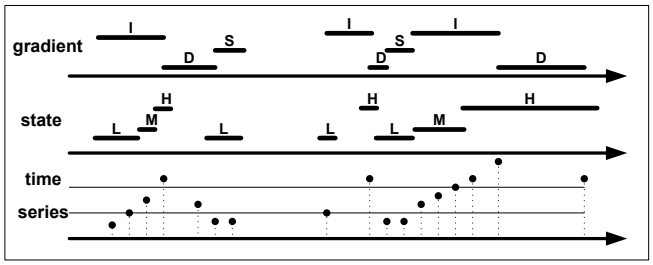


Fig 1. A series of raw time-stamped data of one concept type (at the bottom) is abstracted into an interval-based *state* abstraction (a value classification) that has three discrete values: Low (L), Medium (M), and High (H) (in the middle); and into a *gradient* abstraction (the sign of the first derivative) that has the values Increasing (I), Decreasing (D), and Stable (S) (at the top).

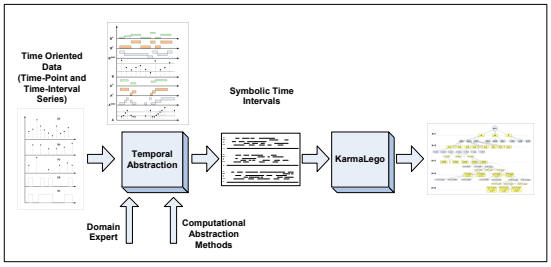


Fig 2. The raw-data time-point and time-interval series are abstracted into a uniform format of [abstract] symbolic time intervals using domain knowledge or specialized computational means. The resulting symbolic time intervals are then mined and a tree of enumerated, sufficiently frequent temporal patterns is generated.

Methods and Discussion

After the patients' data was represented by symbolic time intervals we use KarmaLego to mine and discover all the frequent Time Intervals Related Patterns (TIRPs) [3,4]. Each frequent TIRP represents a cluster of patients that behave in a similar way along time. Eventually, based on this cluster of patients' physiological and biological data a mean centroid is calculated to characterize the group (cluster) of patients, which will enable later to predict new patients behavior along time (i.e., reaction to a drug). In this research we aim at exploring the potential of learning temporal prediction models, and prediction based on a-temporal properties.

References

1. R. Bellazzi, M. Diomidous, I.N. Sarkar, K. Takabayashi, A. Ziegler, A.T. McCray, Data analysis and data mining: current issues in biomedical informatics, Methods of Information in Medicine, 50(6):536-44, 2011.
2. G. Hripcsak, D. Albers, Next-Generation Phenotyping of Electronic Health Records, Journal of American Medical Informatics Association, 20: 117-121, 2013.
3. R. Moskovitch, Y. Shahar, Medical Temporal-Knowledge Discovery via Temporal Abstraction, AMIA 2009, San Francisco, USA, 2009.
4. R. Moskovitch, Y. Shahar, Fast Time Intervals Mining using the Transitivity of Temporal Relations, Knowledge and Information Systems, 2013.

Visualizing Multiple Types of Genomic Information Across Chromosomes With PhenoGram

S.A. Pendergrass, D.J. Wolfe, S. Dudek, M.D. Ritchie

Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Pennsylvania State University, University Park, PA

With the abundance of information and analysis results being collected for genetic loci, user-friendly and flexible data visualization approaches can inform and improve the analysis and dissemination of these data. An ideogram is a graphic representation of chromosomes, and these plots have been used with the addition of overlaid points, lines, and/or shapes, to provide summary information of various kinds coupled with genomic location information. For instance, the results of multiple published genome-wide associations have been plotted on ideograms (www.genome.gov/gwastudies), where lines identify chromosomal location and colored circles indicate associations with different phenotypes/traits. We have developed a flexible software tool PhenoGram, which exists as a web-based tool and also a command-line program, providing a way to visualize data in multiple ways via ideograms. Initially conceived as a method to highlight SNP-phenotype association results across the genome through the use of color-coded circles linked by lines to genomic locations like the NHGRI GWAS catalog plots, the software has been expanded with many unique features. With Phenogram users can plot lines at specific base-pair locations, or base-pair to base-pair regions of chromosomes, with the use of color, with or without other annotation. This feature has been used in several ways to highlight locations or regions of interest. For example, we have used Phenogram to produce plots showing the genomic coverage of SNPs from a genotyping array, plots highlighting the chromosomal coverage of imputed SNPs, as well as plots showing the location of sequenced loci. We have also successfully used this software to plot regions covered by two different copy-number variant detection methods, visually contrasting overlapping regions between the two approaches. Phenogram allows users to annotate chromosomal locations and/or regions with shapes in different colors. This feature can be used to indicate different traits associated with specific loci, and users can choose different shapes to highlight ancestry or another study attributes related to specific data points. Further, users can annotate an ideogram with gene identifiers or other text, as well as create plots showing zoomed in chromosomal locations. PhenoGram is a versatile software, further fostering the exploration and sharing of genomic information. For full details, see http://visualization.ritchielab.psu.edu.

Implementation of Genotype-Tailored Antiplatelet Therapy Following Percutaneous Coronary Stent Placement: A Mixed Methods Study

Josh F. Peterson, MD, MPH; Julie R. Field, PhD; Kim Unertl, PhD; Jonathan Schildcrout, PhD; Daniel C. Johnson, PharmD; Yaping Shi, MS; Ioana Danciu, MS; John H. Cleator, MD; Daniel Johnson, PharmD; Josh C. Denny, MD, MS; Michael Laposata, MD, PhD; Dan M. Roden, MD; Kevin Johnson, MD, MS

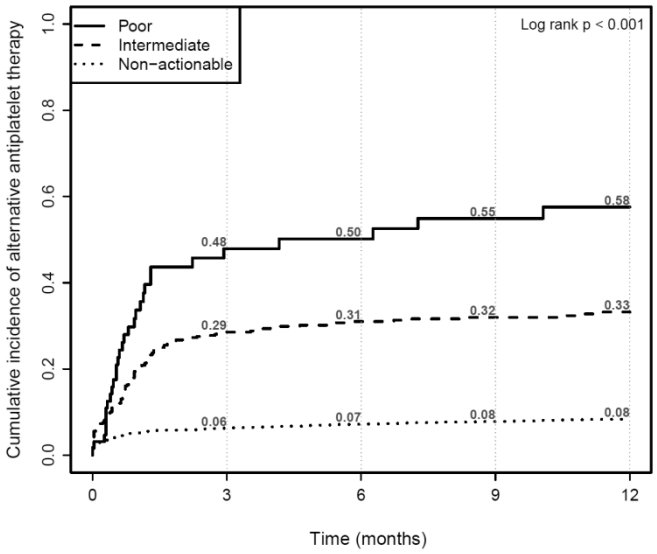
Vanderbilt University Medical Center, Nashville, TN

Summary: Large-scale pharmacogenomic testing of more than 14,000 patients has been implemented at VUMC over the last 3 years in order to tailor drug treatment to genomic variation. We conducted a mixed-methods study incorporating EHR data and semi-structured interviews to assess provider adoption and barriers to genotype-guided antiplatelet therapy in a post-stent population.

Introduction and Background: Genotype-tailored antiplatelet therapy accounting for loss-of-function CYP2C19 variants may prevent ineffective antiplatelet therapy in patients undergoing coronary stenting, but this pharmacogenomic strategy has not been widely implemented or studied. The Pharmacogenomic Resource for Enhanced Decisions in Care and Treatment (PREDICT) program, launched in September 2010, genotypes patients with a multiplexed ADME panel prior to coronary stent placement and delivers clinical decision support (CDS). In this prospective mixed methods implementation study, we sought to determine the impact of the program on selection of antiplatelet prophylaxis for stent thrombosis and physician response to CDS.

Methods: Between 2010 and 2013, interventional and general cardiologists caring for patients receiving a bare metal or drug eluting coronary stent were notified of CYP2C19 variant status via the electronic health record (EHR), computerized CDS, and via a pharmacist-led surveillance program (SP). Patients were followed for 12 months using EHRs to track changes in antiplatelet therapy. Cardiologists were interviewed to determine the facilitators and barriers to prescribing genotype guided therapy. Interviews were transcribed and analyzed using Grounded Theory.

Results: Of 2676 genotyped patients with a coronary stent, 514 (19.2%) were found to have an intermediate or poor metabolizer phenotype. During 12 months of follow-up, 58% of poor metabolizer, 33% of intermediate metabolizer, and 8% of non-actionable CYP2C19 drug metabolism phenotypes were prescribed an alternative to clopidogrel (Kaplan- Meier estimates; Figure). Alternatives consisted of 88% prasugrel, 12% ticagrelor, and 1% high-dose clopidogrel. Rates of genotype tailored therapy were significantly higher among patients without a relative contraindication to prasugrel (41% vs 21%, $p<0.001$) and when a genotype was measured preemptively and available at the time of initiation of therapy (39% vs. 29%, $p<0.001$). Pharmacist surveillance led to more frequent interventions within the 30 days following stent placement than did inpatient or outpatient e-prescribing CDS. After two years' experience with the PREDICT implementation, cardiologists agreed that genomic data could be used to tailor antiplatelet therapy. However, they cited cost of alternative therapy, hand-off of genomic data to outside physicians, complexity of the genomic representations, and gaps in the underlying scientific evidence as barriers to use.



Discussion: When presented with genomic information and CDS, clinicians frequently, but not universally, tailored antiplatelet therapy to CYP2C19 genotype. Tailoring was more frequent when the patient's genotype was measured pre-emptively and meaningful results were delivered in the context of a brief consultation from a pharmacy surveillance team.

Standardized Representation for Electronic Health Record-Driven Phenotypes

Rachel L. Richesson, PhD[1], Shelley A. Rusincovitch[2], Michelle M. Smerek[3], and Jyotishman Pathak, PhD[4]

[1]Duke Univ. School of Nursing, [2]Duke Health Technology Solutions, Duke Univ. Health System, [3]Duke Clinical Research Institute, Durham, NC; [4]Mayo Clinic, Rochester, MN

Summary

Standardized approaches for querying EHR data for research purposes are badly needed. Query requirements of federally-funded projects can be assimilated into a generic computable phenotype template, which includes representation of data and logical query specifications, to support multiple uses including GWAS, personalized medicine, observational research, and pragmatic prospective interventional trials.

Introduction

There is a pressing need for a standard approach to querying EHR data for patient populations with particular phenotypes. Standardized phenotype definitions can support the development of new multi-site studies and ensure comparability of EHR-derived datasets. We assimilate experience from important federally-funded initiatives engaged in EHR-driven phenotyping to identify generalizable information and metadata that are essential for a generalizable representation of phenotype definitions that can be used in heterogeneous healthcare organizations.

Background

Developing standard phenotype definitions for various diseases and conditions is challenging due to the heterogeneity of clinical information systems and the breadth of research topics and data needs. Ongoing experience from NIH-funded initiatives leveraging EHRs for research is shaping a methodology for scalable and valid EHR-driven phenotyping. The Electronic Medical Records and Genomics (eMERGE) Network, funded by the NHGRI, is advancing the use of EHR systems for high-throughput genetics research. The Health Care Systems Research "Collaboratory", supported by the NIH Common Fund, has a broader motivation for the use of EHRs to support pragmatic clinical trials.

Methods

Authors affiliated with the Collaboratory and eMERGE are assimilating experiences and requirements and are developing an EHR phenotype representation model that can be used to represent phenotype definitions for distribution and evaluation. This representation builds largely off of the design of The Phenotype Portal (http://phenotypeportal.org), a tool funded by the SHARPn Project from the Office of the National Coordinator (ONC). Related work from multiple organizations is reviewed and leveraged. For example, the National Quality Forum data quality model and CMS e-measures are designed for use in EHR data and include specific definitions for identifying populations with particular conditions or treatments. Similarly, other relevant work from ONC, HL7, and others are informing a generic phenotype representation model that can support multiple research uses.

Results

Based upon our assimilation of existing efforts, we propose a phenotype representation model for computable phenotypes. The model includes specifications of data elements, data sources (e.g., CMS claims data, EHR clinical data, medication orders), logical operators, and descriptive metadata required to describe phenotypes using data and codes typically collected in EHRs. In addition, specific logical operators and exclusion criteria should be explicitly represented in a computable phenotype template. To support the identification and re-use of computable phenotype definitions, metadata related to the author, endorsements, and versioning are important. Further, the storage, retrieval, and assessment of computable phenotype definitions will be supported by a classification of phenotypes by disease characteristics (e.g., chronic, acute, transient), state of diagnostics (i.e., quantitative measures and indicators of disease exist), and the intended purpose of any computable phenotype definition (e.g., disease management, public health surveillance, quality measurement, observational research, interventional research). Ideally, a standard template for representing computable phenotypes will also include description of details and results from formal validation testing using common metrics (sensitivity, specificity, PPV) in different populations to support evaluation and selection of phenotype definitions by prospective users.

Discussion

Drawing on pioneering efforts and success in the field by eMERGE and others, generalizable aspects of phenotype representation have been identified to support an EHR phenotype representation model which includes data and logical query specifications to support varying uses cases including GWAS, personalized medicine, and pragmatic prospective interventional trials.

Acknowledgments

This publication was made possible by grants 5 U54 AT007748-02 (The Collaboratory), NIH U01HG006379 (eMERGE) and NIH R01GM105688 (PheMA) from the National Institutes of Health.

Biofilter 2.0 for Advanced Predictive Model Development, Testing, and Hypothesis Generation using Expert Domain Knowledge Resources

M.D. Ritchie[1], Alex Frase[1], John Wallace[1], S.A. Pendergrass[1],

[1]Center for Systems Genomics, Department of Biochemistry and Molecular Biology, The Huck Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA, USA

Leveraging the incredible diversity and wealth of biological information collected from the genome, transcriptome, proteome, and other –omic sources, is key for advanced and directed predictive model for the dissection of complex disease. We developed Biofilter for using biological information from public databases to direct complex risk model generation. The first application of Biofilter used 7 sources of domain knowledge: the National Center for Biotechnology (NCBI) dbSNP and gene Entrez database information, Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Gene Ontology (GO), Protein families database (Pfam), and NetPath - signal transduction pathways. Originally Biofilter was for use with gene or single-nucleotide polymorphism (SNP) based data, such as filtering genome wide association (GWA) genotypic data for biologically directed model generation. For example, through directing the search space using Biofilter with GWA data, pairwise SNP-SNP interaction models were developed, replicating across datasets. We expanded and updated Biofilter to version 2.0 in several important ways. First, we migrated the domain knowledge used by Biofilter into a separate database called the Library of Knowledge Integration (LOKI) and incorporated 4 additional data sources: Molecular INTeraction database (MINT), Biological General Repository for Interaction Datasets (BioGrid), Pharmacogenomics Knowledge Base (PharmGKB), Open Regulatory Annotation (ORegAnno), and the database of Evolutionary Conserved Regions (ECRBase). The independence of LOKI from Biofilter 2.0 facilitates easy update of the data sources contained within LOKI, reflecting the update of public LOKI data sources. Biofilter 2.0 now accepts a greater range of data types, including SNP, rare variant, copy-number variation (CNV), and evolutionary conserved region (ECR) data. Biofilter 2.0 allows more user customization of filtering and annotation, and easier use of custom data sources, providing a flexible framework for the integration and user-driven combination of multiple data types, providing a way to use the ever-expanding expert biological knowledge that exists to direct complex predictive models for elucidating the etiology of complex phenotypic outcomes. The software is available for use at http://ritchielab.psu.edu/.

We will demonstrate the utility of Biofilter 2.0 using several natural biological datasets involving genomic data integrated with electronic health records.

SPIRIT – Integrated Platform for Protocol Decision Trees, Eligibility Screening and Cohort Identification

Ajay Shah PhD, John Meng MD, MS, Sai Achuthan PhD, Srinivas Bolisetty MS, MTech, Ayyappan Nagender, Joyce C. Niland PhD, City of Hope, Duarte, CA

Abstract
SPIRIT an informatics platform seeks to integrate basic, clinical and translational research applications. It provides graphical representation of differential eligibility criteria for clinical trials and an expert system to identify an applicable clinical trial. Integration with i2b2 allows scanning EMR system for additional eligible cohorts and analysis using machine learning.

Introduction and Background
Clinical protocol portfolio management, cohort identification and cohort stratification pose significant challenges in Clinical and translational research. Generally, there are competing clinical trials vying for patients. Multiple incoherent tools pose data integrity and usability challenges. A platform approach to solving this problem allows us to comprehensively address this problem.

Methodology
SPIRIT is being implemented as an integrative n-tier software platform for building and integrating informatics applications using standard software component. SPIRIT enables clinical and translational research document management, decision support, natural language processing, computational chemistry and biology. The foundational pieces of the platform include Oracle's FUSION SOA Suite, Pipeline Pilot, statistical package R, expert system CLIPS (C Language Integrated Production System) etc.

Decision Tree Tool (SPIRIT DT) provides functionality to visually represent eligibility criteria and inspect, monitor and analyze clinical trials portfolio. A decision tree is converted to rules representation used in SPIRIT Expert System (SPIRIT ES), a customized CLIPS interface. Any change in protocol decision tree is dynamically reflected in SPIRIT ES. SPIRIT ES provides wizard driven search for a clinical trial, eligibility criteria where no trials exist and rules based optimum match for clinical trials.

SPIRIT ES interfaces with EMR data in enterprise data warehouse (EDW) via i2b2. The eligibility criteria are converted to an i2b2 query, to search for additional cohorts. Currently, the number of cohorts returned is a superset of eligible cohorts because not all data needed for protocol eligibility criteria is available in EDW. Potential cohorts available in EMR are displayed in Decision Trees. SPIRIT Machine Learning (SPIRIT ML) incorporates clustering to provide further insights into cohort characteristics

Results and Discussion
We applied SPIRIT ES to Hodgkin Lymphoma (HL) protocol decision tree. There are five clinical trials for HL at City of Hope. Table 1 shows the current and expected enrollment in the clinical trials. The last column shows the number of patients matching the eligibility criteria in the EMR system who may be approached for enrollment. Clustering of these eligible patients using SPIRIT ML provided further insights into the patient characteristics.

Table 1:

Protocol	Current Enrollment	Target Enrollment	Potential Eligible Patients in EMR
1	2	10	7
2	1	6	8
3	12	33	34
4	19	37	141
5	2	37	73

Pharmacogenomic Analysis of Pathways in Multiple Rat Strains: Implications for Drug Testing in Models

Mary Shimoyama PhD, G. Thomas Hayman PhD, Stan Laulederkind PhD, Rajni Nigam MS, Victoria Petri PhD , Jennifer R. Smith MS, Shur-Jen Wang PhD, Jeff De Pons BS, Pushkala Jayaraman MS, Weisong Liu PhD, Marek, Tutaj MS, Elizabeth Worthey PhD, Melinda R. Dwinell PhD, Howard Jacob PhD

Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, Wisconsin

Abstract The rat is an important model used to study effects of drugs. Whole genome sequencing provides a rich source of data concerning potentially damaging variants within genes interacting with drugs. Identifying variants of potential significance for individual rat strains can provide researchers with additional information when choosing models for their studies.

Introduction The laboratory rat has been an important model for studies in disease mechanisms, general physiology, pharmacology, and toxicology and strains exhibiting specific phenotypes have been developed for use in these studies. Investigators have been dependent on strain specific phenotype profiles in choosing appropriate models, but with whole genome sequencing investigators can now add genomic profiles as well. These can include pharmacogenomics profiles for individual strains and comparisons across strains to assist investigators in choosing models for drug studies.

Methods Variation data for 40 rat strains available at the Rat Genome Database (RGD) (http://rgd.mcw.edu) were analyzed to determine non-synonymous status, and using PolyPhen, the likelihood they are benign or damaging. The genes in which the non-synonymous variations lie were analyzed using the Gene Annotator Tool (http://rgd.mcw.edu/rgdweb/ga/start.jsp) for known interactions with drugs and chemicals based on annotations at RGD imported from the Comparative Toxicogenomics Database (http://ctdbase.org/) and participation of their human orthologs in known pharmacodynamics and pharmacokinetic pathways illustrated at PharmGKB, (http://www.pharmgkb.org/). Variant profiles for each strain for genes identified in specific pharmacodyamic and pharmacokinetic pathways were created as well as for all other genes identified as interacting with drugs. Genes associated with specific drug classes such as anti-hypertensive drugs, anti-neoplastic agents, anti-inflammatory drugs, and immunosuppressants were also analyzed for each strain to identify patterns of variants.

Conclusion In some cases, such as with Warfarin, no damaging variants were found in any of the strains analyzed indicating multiple strains can be used for drug testing. Similarities and differences in variations in genes associated with a variety of drugs were found among substrains as well as across parental strains. Specific, limited, potentially damaging variations in a few strains for important genes were found as well as multiple, widely varying SNVs across all strains in other gene subsets indicating these likely have little impact. In some cases, genes that are part of the human pharmacodyanmic or kinetic pathways are represented by two homologs in rat which could affect drug testing in rat. In some cases, there were two receptors in rat with few damaging variants found in one and potentially damaging variants were found in multiple strains in the other receptor, indicating more research is needed to determine what effect this has on drug efficacy in rat as compared to human. Some strains showed distinct patterns of variation across gene sets while others showed few patterns. While typically, drug testing is done on a single rat strain deemed to be “normal” or on a strain exhibiting the specific phenotype of interest, examining variant analyses could prove valuable in choosing rat models for drug testing by allowing researchers to choose strains without potentially damaging variants or choosing a panel of strains with differing variants that could more closely represent the potential human population variations.

AACT-Results: The Results dataset extensions for the AACT database

Asba Tasneem, PhD[1], Skip Maza[1], Surendra Gonigunta[1], James Topping[1], Philip D'Almada[1], Karen Chiswell, PhD[1], Robert Califf, MD[2]

[1]Duke Clinical Research Institute, [2]Duke Translational Medicine Institute, Durham, NC

Abstract

An analysis dataset is required to permit aggregate analysis of the ClinicalTrials.gov dataset. A publicly available registry dataset that includes study registration and results fields was imported and reformatted into a relational database that facilitates aggregate analysis.

Introduction

ClinicalTrials.gov, the world's largest repository of information on clinical research studies, houses data from more than 150,000 studies conducted in more than 180 countries. Although registry data are available for bulk download, issues related to data structure, nomenclature, and changes in data collection over time limit the quantitative interpretation of these data.

Background

With support from the Clinical Trials Transformation Initiative (CTTI), we built a relational database comprising information from study registration fields downloaded from ClinicalTrials.gov—the database for Aggregate Analysis of ClinicalTrials.gov (AACT) (Tasneem et al., 2010). In the present study, we report on our expansion of the AACT database to include primary trial outcomes and adverse events reporting fields (Results dataset) from ClinicalTrials.gov.

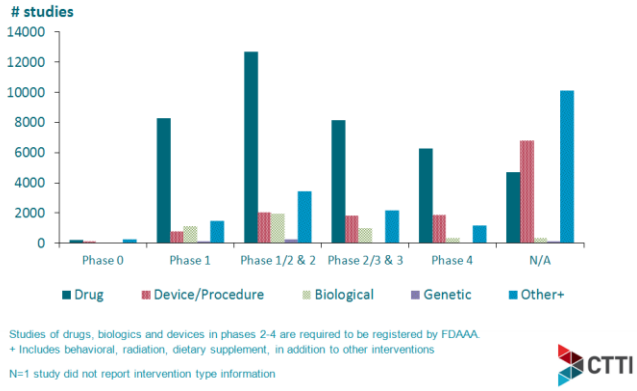
Methods

A dataset encompassing 133,128 clinical trials was downloaded from ClinicalTrials.gov on September 27, 2012 in XML format. ClinicalTrials.gov's XSD was used for initial design and data modeling of new Results fields in the dataset. Informatica Power Center ETL process (Extract, Transform, Load) was used to load data into Oracle tables. Internal checks were performed in landing and staging tables to validate cardinality of data. Tidal Enterprise Scheduler was used to schedule automatic updates of the AACT database. The pipe-delimited text files, SAS CPORT transport files, and Oracle dump files were created as a resulting output.

Results

We created an analysis dataset from ClinicalTrials.gov that can be used for aggregate analysis, facilitating the use of data by researchers seeking to study and contrast characteristics of specific groups of trials. Trial characteristics can be summarized by design (interventional vs. non-interventional), sponsorship (NIH/Industry/Other), geographical location, phase, intervention type [Fig. 1], age-based eligibility criteria, number of study arms, and various other parameters. About 5% (7,003/133,128) of studies reported baseline characteristics, outcomes, and adverse event data.

Fig 1: Study Registration by Intervention Type and Phase



Discussion

The AACT database affords researchers a unique opportunity to examine the ClinicalTrials.gov dataset using one of the three new available formats at an aggregate level. Researchers can use outcomes and adverse events summary data reported at ClinicalTrials.gov to conduct meta-analyses or systematic reviews (e.g., to compare the efficacy and safety of different types of diabetes therapies). However, only a small subset of studies currently available in ClinicalTrials.gov were registered after rules mandating expanded reporting were enacted, thereby limiting the amount of data available for comparison. Therefore, for the immediate future, the Results dataset from ClinicalTrials.gov will most likely be a useful supplement to traditional data sources used for meta-analysis or systematic review, such as published and unpublished manuscripts and abstracts.

Acknowledgments

Financial support for this work was provided by the U.S. Food and Drug Administration grant U19 FD003800 awarded to Duke University for CTTI.

Identifying Patients with Hypertension in the Electronic Medical Record

Pedro L. Teixeira; Wei-Qi Wei MMed, PhD; Robert M. Cronin, MD;
Joshua C. Denny, MD, MS
Vanderbilt University, Nashville, TN

Abstract: *Hypertension is one of the most common primary diagnoses – affecting one third of Americans. It is estimated to contribute to one in six adult deaths in the US, and is an important covariate for many analyses. Yet there is no computational algorithm available to phenotype individuals based on electronic medical record (EMR) data. We have evaluated the accuracy of ICD9 codes, medications, blood pressure, and textual mentions of "hypertension" for identifying hypertensive individuals. Individually, medications yield an area under the receiver operator characteristic curve (AUC) 0.886. The best results are achieved with algorithms that combine inputs – raw sum of multiple element counts and logistic regression across elements with AUCs of 0.919 and 0.908.*

Introduction: The seventh report from the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure defines hypertension as a consistent blood pressure over 139 mmHg systolic and/or 89 mmHg diastolic. It is a key modifiable risk factor for cardiovascular disease. Identifying patients with hypertension based on EMR data is challenging due to irregularity and incompleteness. Data is often not recorded, and individuals controlling hypertension via lifestyle modification are often not documented. Furthermore, blood pressure readings above the threshold used to diagnose hypertension are often caused by a variety of temporary sources including – e.g., acute distress, adverse reactions, psychological stress. We examined the sensitivity and specificity of various individual data categories as well as expert-designed algorithms leveraging combinations of the aforementioned data categories. Our simple algorithms perform well in spite of the untamed nature of EMR data.

Methods: We used the deidentified EMRs of a random subset of 303 individuals with "regular outpatient care" at Vanderbilt – defined as at least two outpatient visits and two vitals readings between 1/1/07 -1/1/09. Records were reviewed by two physicians and a medical student with 20% overlap. Conflicting classifications were arbitrated by a fourth physician. Counts were determined for hypertension-related ICD9 codes, medications indicated to treat the condition, hypertensive blood pressures, and mentions of hypertension (searching for "hypertension"/"HTN" within problem lists, history and physical notes, and discharge summaries excluding "pulmonary [arterial] hypertension"). Medications were identified using MedEx. Multiple occurrences within a single day were counted as one.

		(Present/Absent)	
Inputs	**AUC (95% C.I.)**	**Sensitivity**	**PPV**
ICD9	0.874 (0.839-0.910)	0.794	0.946
Drugs with HTN Indication	0.886 (0.843-0.929)	0.935	0.823
High Systolic	0.882 (0.836-0.928)	0.965	0.803
High Diastolic	0.734 (0.677-0.791)	0.744	0.809
High Systolic OR Diastolic	0.875 (0.828-0.922)	0.970	0.794
PL+DS+HP Mentions	0.872 (0.829-0.916)	0.920	0.813
Raw Sum of 1, 2, 5, and 6	0.919 (0.881-0.957)	1.000	0.713
1 of 4 elements	0.570 (0.535-0.605)	1.000	0.713
3 of 4 elements	0.867 (0.824-0.910)	0.884	0.926
Logistic Regression 5fold CV	0.908 (0.827-0.988)	0.980	0.680

Results: Of the 303 patients examined, 292 had sufficient information in their EMR to classify each as case or control. Interreviwer agreement was high, with a kappa of 0.92. Results for the various approaches, including confidence intervals about the area under the receiver operator curve (AUC, calculated using the DeLong method), are given in the table. Diastolic measures alone resulted in the lowest accuracy. All other elements result in an AUC > 0.87. Medications perform best of the individual items (AUC 0.886). Accepting any element of the four results in the lowest performance (AUC 0.570). However, using a sum of the counts from ICD9 codes, medications, high BP, and hypertension mentions and logistic regression perform best with AUCs of 0.919 and 0.908 respectively.

Discussion: We have shown that medications with a HTN indication are among the most informative data elements for identifying hypertensive individuals within the EMR. In addition, simple algorithms combining elements result in the highest accuracy with AUC surpassing 0.91. Since HTN is an important covariate in many clinical and genomic studies, in which both recall and precision are important, these results provide guidance on methods to approach identification of HTN patients. We believe further refinement of concept identification and aggregate statistics based on the available data will enable further performance improvement. Furthermore, we will test our algorithm at other sites to assess portability.

User Requirement Analysis for the database of Genotypes and Phenotypes (dbGaP): A Multidimensional Approach for Query Tool Design

Rebecca Walker, BS, Hyeoneui Kim, PhD, Stephanie Feupe Feudjio, MS, Seena Farzaneh, MS, Mindy Ross, MD, Son Doan, PhD, Lucila Ohno-Machado, PhD, Ko-Wei Lin, PhD
Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA

Abstract: *In order to improve usability of dbGaP and to develop PhenDisco, a phenotype query tool, we conducted user requirement analysis using three approaches: online survey with broad range of users, in person interview with dbGaP users, and data request analysis. The results provided us the precise description of the current content, functionality, and quality needs for dbGaP users, and gave valuable feedback for PhenDisco development.*

Introduction and Background

dbGaP (http://www.ncbi.nlm.nih.gov/gap) is a public database which stores various phenotypic and genotypic data of genome-wide association studies (GWAS). Reusing these datasets can promote efficiency in research and scientific discovery. However, inaccurate query results make current dbGaP lack usability. In order to provide users more accurate retrieval of phenotypic data in dbGaP, Phenotype Discoverer (PhenDisco, http://pfindr.net), a web-based query tool, has been developed at University of California San Diego. Understanding how to improve dbGaP and developing a tool to fit the needs of users are both critical during system development. In this study, we performed three types of user requirement analysis. First, we developed an online survey questionnaire to reach a broad range of users and asked users about their experience interacting with the dbGaP website. Secondly, we conducted in person interviews with local dbGaP users. Lastly, we analyzed data request records and categorized the information. The overall goal of this project is to determine the needs and conditions to meet in PhenDisco design.

Methods

Online Survey: We developed an online survey of user experience with 16 questions. The content of the questions were based around ease of use and satisfaction with results of searches. The survey design consisted of unstructured, open-ended questions as well as structured, Likert scale questions. The open-ended questions allowed users to input reasons for satisfaction or dissatisfaction and make suggestions. The structured questions were designed to make the survey quick and of little effort to the users and to allow users to rank different aspects (satisfaction, helpfulness, ease, adequacy) of the tools available on the dbGaP website. We sent out 200 email invitations twice, using a mass email approach, to the online survey to the users who have requested data from dbGaP.

In Person Interviews: We conducted 8 interviews with local dbGaP users following specific face-to-face interview guidelines. The interview touched both on usability and use case aspects. The guidelines allowed the interviewees to provide cases they would use in their research and to suggest new functions they would be interested in using.

Data Request Analysis: We analyzed 14,287 publically available data use requests from dbGaP. The data use requests were pre-processed with a text-processing pipeline and then run through MetaMap to map the pre-processed text into UMLS concepts. The output consists of standardized semantic types, Concept Unique Identifiers (CUIs), and concept names. These were then manually reviewed and characterized by domain experts.

Results and Discussion

We analyzed 17 complete responses from 29 total responses from online survey. More users preferred to use just the basic search tool, as opposed to the advanced search option, but most users were satisfied with the ease of use with both search tools. We found dissatisfaction with the ease of review of retrieved studies as well as inaccuracy and incompleteness of the results. Open-ended questions pinpointed specific items to address during developing new tool; these topics include implementation of standard terminology, highlighting key search terms in results, and the ability to compare studies. For the in person interviews, most suggestions, such as keyword highlighting, overlapped with those from the online survey. Other feedback includes ranking of results by relevancy, presenting more metadata in the results summary page, the need for term standardization and the expansion of search terms by synonyms and hierarchical concepts. The data request analysis identified the most frequent topics searched for in dbGaP: Diseases(30%), Test and/or Procedures(13%), and Chemical or Biological Substance(8%). In the Disease category we found the highest requested topics are Neoplasms/Cancer (30%), Psychiatric Disease (13%), Genetic Disease Congenital Abnormality (8.6%), and Cardiovascular Disease (8.1%). The results provided the user requirements for PhenDisco development. We implemented most of the feedback in the final version of PhenDisco and provided the user with a more satisfied experience.

Acknowledgement: This project was supported by grants UH2HL108785 and U54HL108460 (NIH/NHLBI).

Cancer Patient Integrative Stratification via a Two-step Consensus Clustering of Molecular Expression and Clinical Attributes

Chao Wang, MS[1], Raghu Machiraju, PhD[2], Kun Huang, PhD[1]
[1] Department of Biomedical Informatics, The Ohio State University
[2] Department of Computer Science and Engineering, The Ohio State University

Abstract
In recent decades, many studies have led to biomarkers for cancer outcome predictions, which assist clinicians on selecting the right treatment strategy. These biomarkers include both pathological attributes and various types of omic data. However, there is a lack of a unified means for patient stratification which can effectively integrate the heterogeneous types of molecular and clinical data and improve accuracy on patient outcome prediction. In this paper, we propose a novel two-step cancer patient stratification workflow, which aggregates clinical information and molecular expression profiles. First, stratification based on gene and miRNA expression profiles is achieved by an integrative clustering approach. Then, the task of comprehensive patient subtyping based on both clinical attributes and the molecular stratification is formulated and solved as a consensus-clustering problem. The results obtained from the TCGA breast cancer study suggest that this approach out-performs stratification based on any single clinical attributes or molecular data type.

1 Introduction

In recent years, the advanced researchers have brought personalized medicine into a brand new level by introducing technologies such as DNA microarray, mRNA sequencing and miRNA sequencing[1 2]. These genome-wide studies have heralded new era in clinical practice to account for variations across population. Risky groups classified by the status of the biomarkers are given more aggressive treatments, and prediction of cancer growth, proliferation and metastasis is also based on these 'omic-generated' biomarkers. Specifically, many types of cancer can be reclassified into distinct subtypes based on the genomic makeups and molecular profiles and these subtypes present different clinical outcomes including prognosis and response to treatments [3 4 5]. An example of success is the PAM50 gene panel being applied to breast cancer patient stratification[4].

Although the recent-developed high-through-put data have associated subtypes of cancers to these prognostic and predictive biomarkers, this data-driven approach of uncovering gene signatures is often hard to interpret due to the knowledge limitation about the biological functions of these biomarkers[6]. In addition, traditional stratifications of patients based on cancer staging, histological subtyping and other pathological clinical attributes are still the essential diagnosis means in clinical practice. However, these clinical knowledge-based classifications do not necessarily agree with the molecular stratifications. Furthermore, the multiple labels for each patient can be confusing and difficult to interpret. With all available patient stratification methods, there is a need for a consensus classification of patients implemented by both clinical information and molecular biomarkers.

Therefore, in this paper we aim to develop a novel method that can integrate information of both clinical attributes and molecular data to provide a consensus clustering for patients. We focus on two major challenges on using ensemble of molecular expression data and clinical attributes: (1) Integrative clustering of the multiple molecular expression data; (2) Aggregation of all base-clustering labels. Specifically, we propose a novel two-step consensus clustering workflow integrating both the molecular clustering and clinical attributes. This workflow resolves the first challenge using a Canonical Correlation Analysis (CCA) based method to cluster the multi-types of data. Then it addresses the second challenge with a Bayesian consensus clustering method. The results obtained from applying this approach in The Cancer Genome Atlas (TCGA) breast cancer data demonstrated the power for this approach in differentiating the patient population with different outcome and its potential in integrative genomics studies.

2 Methods and Materials

2.1 TCGA breast cancer dataset and preprocessing

TCGA is a project aiming to demonstrate genetic mutations for various types of cancers using multiple genomic data[7]. mRNA and miRNA expression profiles were collected from 441 primary tumors of breast cancer patients who have all mRNA, miRNA and clinical data. The median follow-up is 3 years. Expression profiles of mRNA and miRNA were converted from level-3 mRNAseq and miRNAseq with RPKM (read per kilobase per million) values downloaded from TCGA data portal. The mRNA and miRNA expressions were log2 transformed and normalized. The clinical attributes (disease stage, tumor grade and histological type) were discretized from 1 to the number of

categories. Disease stage determines the extent to which a cancer has developed by spreading. Tumor grade measures cell appearance in tumors and other neoplasms. Histological type categorizes the heterogeneity found in breast cancer based on architectural features and growth patterns.

In order to eliminate the noise in the mRNA expression profiles, we selected 70 prognostic genes and 7 prognostic miRNAs from previous studies[14 15]. Out of the 70 genes, we found 28 genes exactly matched (Table 1) with genes in TCGA data.

Table 1. Prognostic mRNAs and miRNAs used in the integrative molecular clustering.

Data Type	Prognostic Biomarkers
Gene Expression	AKAP2,AP2B1,BBC3,CCNE2,CENPA,COL4A2,DCK,ECT2,ESM1,EXT1,FGF18,FLT1,GMPS,GNAZ,GSTM3,IGFBP5,MCM6,MMP9,NMU,ORC6L,PECI,PRC1,RAB6B,RFC4,SERF1A,SLC2A3,TGFB3,WISP1
miRNA Expression	hsa-mir-1307,hsa-mir-148b,-mir-210,hsa-mir-328,hsa-mir-484,hsa-mir-874,hsa-mir-93

2.2 The Workflow

Instead of manually integrating separate patient clusters, our workflow achieves data integration through consensus clustering which combines all the cluster information as a unified stratification, which is a single partition of patient samples obtained by integrating multiple measurements. Prior to this study, to our knowledge there is no bioinformatics method that can integrate both numerical molecular measurements and categorical attributes. Current cluster ensemble algorithms, such as the cluster-based similarity partitioning algorithm (CSPA)[8], hypergraph partitioning algorithm[8], or Bayesian based algorithms[9] are able to accomplish the integration of clusterings. However, none of them was designed to address the integration with numerical measurements such as gene expression. And, most of the integrative analytical methods require specific data type, but our method can be applied any type of omic data and other clinical attributes of the patients, such as distant metastasis status and lymph node infiltration.

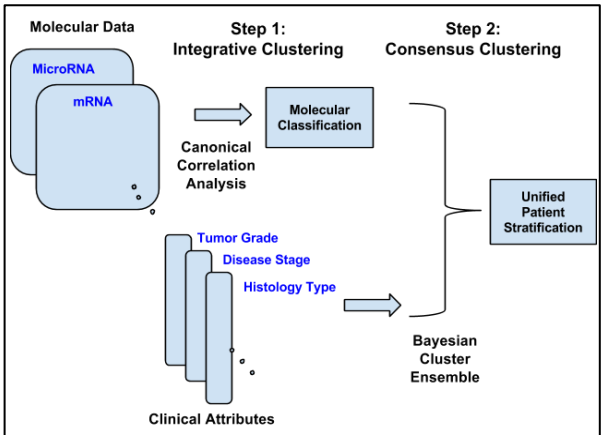


Figure 1. Schematic plot for patient Stratification proposed in this paper.

Figure 1 provides an overview of the framework for the proposed patient classification approach. Specifically, in the first step, the salient microRNAs and mRNAs are selected and the patients are clustered based on the expression of these salient molecules. There are a few integrated unsupervised clustering algorithms developed to fulfill this task[10 11 12]. Considering the relative large sample size in our study, we adopt the CCA-based multiple-view clustering algorithm[11]. Then, in the second step the cluster labels generated from the molecular expression profiles are jointly considered with clinical labels by consensus clustering, and a final label is constructed for each patient. Since it is not clear which measurement of similarity between two distinct clustering with both molecular and clinical data being used, a Bayesian Cluster Ensemble[9] method is selected.

2.2.1 Integrative Molecular Stratification using Canonical Correlation Analysis

In the first step, we have multiple measurements (molecular expressions in this paper) from n modalities, with modality ($j = 1,2,\dots,n$) from a mixture of k Gaussians ($D_1^j,\dots,D_k^j$). The goal of integrative molecular clustering is to recover the k sub-distributions representing distinct molecular subgroups.

In this study, the molecular data are gene expression and miRNA expression, so there are two sources in this setup ($n = 2$). In particular, we assume i-th sample can be represented as $x_i = (x_i^{(1)}, x_i^{(2)})$, where $x_i^{(1)} \in \mathbb{R}^{d1}$, $x_i^{(2)} \in \mathbb{R}^{d2}$. $d1$ and $d2$ are the dimensions of two types of data respectively. And, $\mu_i^{(1)}$ and $\mu_i^{(2)}$ are the means of distribution i in measurement 1 and measurement 2, respectively. In genomic data preprocessing, the expression data is usually normalized and centered at zero mean, so $\mu^{(1)} = \mu^{(2)} = 0$. Hence, we can define the covariance matrix of the measurements $x = (x^{(1)}, x^{(2)})$ as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{21} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} E[x^{(1)}x^{(1)^T}] & E[x^{(2)}x^{(1)^T}] \\ E[x^{(1)}x^{(2)^T}] & E[x^{(2)}x^{(2)^T}] \end{bmatrix}$$

If we assume that the two measurements are independent to each other, which usually holds for miRNA expression and mRNA expression, then for each distribution r in the mixture Gaussian:

$$E\left[x^{(1)}x^{(2)^T}\right] = E\left[x^{(1)}|r=i\right]E\left[x^{(2)^T}|r=i\right] = \sum_i \omega_i E\left[x^{(1)}\right]E\left[x^{(2)}\right]^T = \sum_i \omega_i \mu_i^1 \cdot (\mu_i^2)^T.$$

Previous publications[11 13] showed that the column spans of matrix U and V from reduced SVD of $E\left[x^{(1)}x^{(2)^T}\right] = UDV^*$ are subspaces spanned by the means in measurements 1 and 2, respectively. Therefore, following the method proposed by Chaudhuri et al.[11], the integrative clustering of mRNA and miRNA expression can be described as followed:

1. Partition the samples into two subsets X and Y randomly.
2. Compute the empirical covariance matrix Σ_{XY} between mRNA and miRNA expressions.
3. Compute the singular vectors of $\Sigma_{12}(X)$ ($\Sigma_{12}(Y)$ respectively) and project Y (X) on subspace spanned by the top $k-1$ left singular vectors.
4. On the projected data, apply any single-linkage clustering algorithm.

2.2.2 Obtaining the Final Patient Stratification from Consensus Clustering

With all patients assigned to a certain clustering based on multi-source molecular data, the information of molecular heterogeneity is encoded into the labels of this population. In the second step, let us consider the method to integrate the molecular labels with the clinical attributes. Given M sets of clustering $C_j = \{c_{ij}, [i]_1^N\}$ drawn from N objects $D = \{d_i, [i]_1^N\}$, the cluster label vector for object d_i is denoted by $x_i = \{x_{ij}, [j]_1^M\}$.

The goal is to construct a model that can find the consensus cluster integrating these M sets of clustering. There are many algorithms to solve this problem. General speaking, these methods fall into two categories: probability model and similarity model. The former one postulates a probability model to determine the labels of the objects and solve via maximum likelihood optimization, while the latter one seeks the consensus clustering with the largest concordance (similarity) with original clusterings. The second approach requires a measurement of similarity between two distinct clustering, which is usually not trivial to find, thus we adopt a Bayesian model[9] in this paper.

2.3 Assessing parameterization of consensus clustering

To determine the parameter k, the number of final clusters in consensus clustering, we aim to optimize the similarity between the final clustering and the base clusterings. Here, normalized mutual information function is used to measure the similarity between clusterings. In other words, the parameter $k$can be found, so that the similarity between the final cluster and all based clusterings is maximized.

2.4 Survival Analysis

For each set of patient classification, a low-risk group and a high-risk group were identified as the best separation from combinations of the resulting subgroups. Survival was analyzed according to the Kaplan-Meier method. The differences between survival distributions of the two groups were evaluated by log-rank test. A supervised analysis was performed using ANOVA (false-discovery rate of 5%) followed by Tukey post-hoc testing to identify genes with differential expression between pre-defined groups.

3 Results

We introduce this two-step model to address the challenge of integrative patient subtype identification based on both molecular expression and clinical attributes. In this section, the proposed workflow is applied to identify integrative patient stratification to breast cancer patients in TCGA. In this section, the identified breast cancer patient subgroups were evaluated computationally, clinically and biologically.

3.1 Selection of the Number of Clusters

It is desirable to partition the patient population into groups that can maximize the mutual information between the final partition and the base clusterings. After the integrative

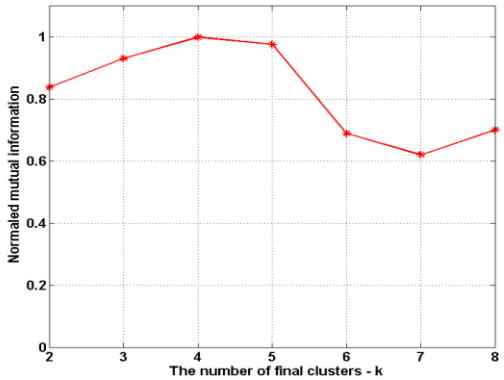


Figure 2. The mutual information of the consensus clustering with original labels under different choices of k.

molecular partition was obtained, different choices of the parameter $k \in [2,8]$ were tested by evaluating the normalized mutual information defined in the above section. We repeat the clustering for $r = 10$ times. In the case when $k = 4$, the final clustering has the largest mutual information with the base clusterings (Figure 2), so in the rest of the paper, we use $k = 4$.

3.2 Integrative Clustering of Multiple Types Molecular Expression Data

In this section, we will show the superiority of using multiple types of molecular expression data over the method of using just one. Based on the 28 genes previous literature has found, the whole TCGA population can be grouped into 4 groups using Kmeans and then finally assembled as two groups by selecting the most differentiating combination

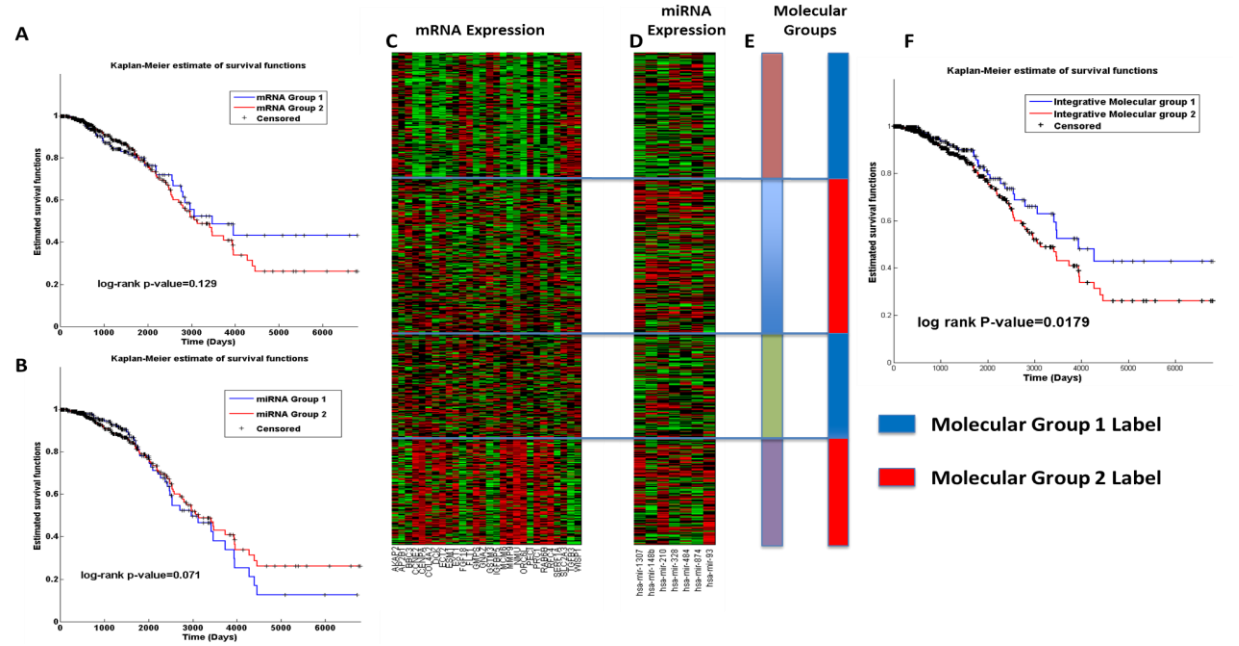


Figure 3. Single type of molecular data clustering based stratification versus Integrative molecular subgrouping based on the 28 prognostic genes and 7 prognostic miRNAs using CCA. (A). Kaplan-Meier curve for the two groups of patients stratified based only mRNA expression. (B). Patient stratification using only miRNA expression. (C). Heatmap for 28-mRNA expression profile after CCA multi-view clustering. (D). Heatmap for 7-miRNA expression profile after CCA multi-view clustering. Each row represents miRNA expression of one patient, which is aligned with mRNA profile in (C). (E). Molecular grouping labels (k=4). The 4 groups were further merged to two groups (labeled as blue and red) based on the best separation of survival curves. (F) Kaplan-Meier plot of the two molecular groups.

(Figure 3A) for comparison purposes. The same method is used to classify breast cancer patient using the existing prognostic miRNA biomarkers (Figure 3B). Although the gene and miRNA signatures selected in the clustering analysis are tested to be powerful in prognosis of breast cancer patient outcomes, neither of them shows significant difference when applied to this data, with P-values of 0.129 and 0.071 respectively.

Integrative clustering takes both mRNA and miRNA inputs and conducts integrative clustering on both of these omic data. Therefore, unlike the single-linkage clustering method, it can measure multiple levels of the characteristics of the complex tumor. Heatmaps in Figure 3C and Figure 3D show the resulting clustered profiles of 28 genes and 7 miRNAs, respectively. Visualized from the heatmaps in Figure 3, each cluster shows unique patterns in both of their mRNA expression and miRNA expression, indicating the efficacy of integrative clustering. The four molecular subpopulations (labeled by Figure 3E) were then combined into two groups, whose Kaplan Meier survival curves were shown in Figure 3F. Molecular group 1 (label in blue) has worse survival rate than molecular group 2 (labeled in red).

Non-coding small RNA is functioning as important molecular involving in many cellular processes. Our method shows that with the assistance of miRNA expression, molecular stratification of TCGA breast cancer patients is boosted upon mRNA expression. In addition, mRNA expression also boosts miRNAs in prognosis.

3.3 Integration of Molecular Expression and Clinical Attributes

Using integrative molecular clustering, we can cluster the cancer patients in two several molecular groups as labeled in Figure 3F. Upon the molecular stratification and the clinical annotations (tumor grades, disease stages and histology types) described earlier in this section, consensus clustering is applied to obtain a unified patient classification.

Disease Stage (DX stage) is a better indicator for patient stratification regarding to survival than clinical attributes, such as tumor grade and histological type. It is not surprising to find that classification based on disease stage (Figure 4A) is more powerful than others (Figure 4B and 4C). However, with all three clinical labels and molecular subtype labels integrated in consensus clustering, the patient population is even better (p=0.00195) differentiated into a low-risk group and a high-risk group (Figure 4D).

3.4 Subtype-specific Genes of High-Risk Group

We observed significant expression differences of 93 genes between the subgroup with poor outcomes and the rest. Toppgene enrichment analysis was done for the list of 93 genes. The list of significantly enriched pathways includes regulation of gene expression in beta cells, cell cycle, DNA damage and Metabolism of proteins, etc. IPA gene network analysis was conducted on these 93 genes and the one with the top score was shown in Figure 4E. These genes include CDH1 and RUNX1, which have been tested for their role in breast cancer. We observed that somatic mutation incidence of RUNX1 in the poor outcome group (18%) is three times larger than that of the rest (5.4%). The somatic mutation incidence of CDH1 in the poor outcome group (29.4%) is significantly larger than that of the

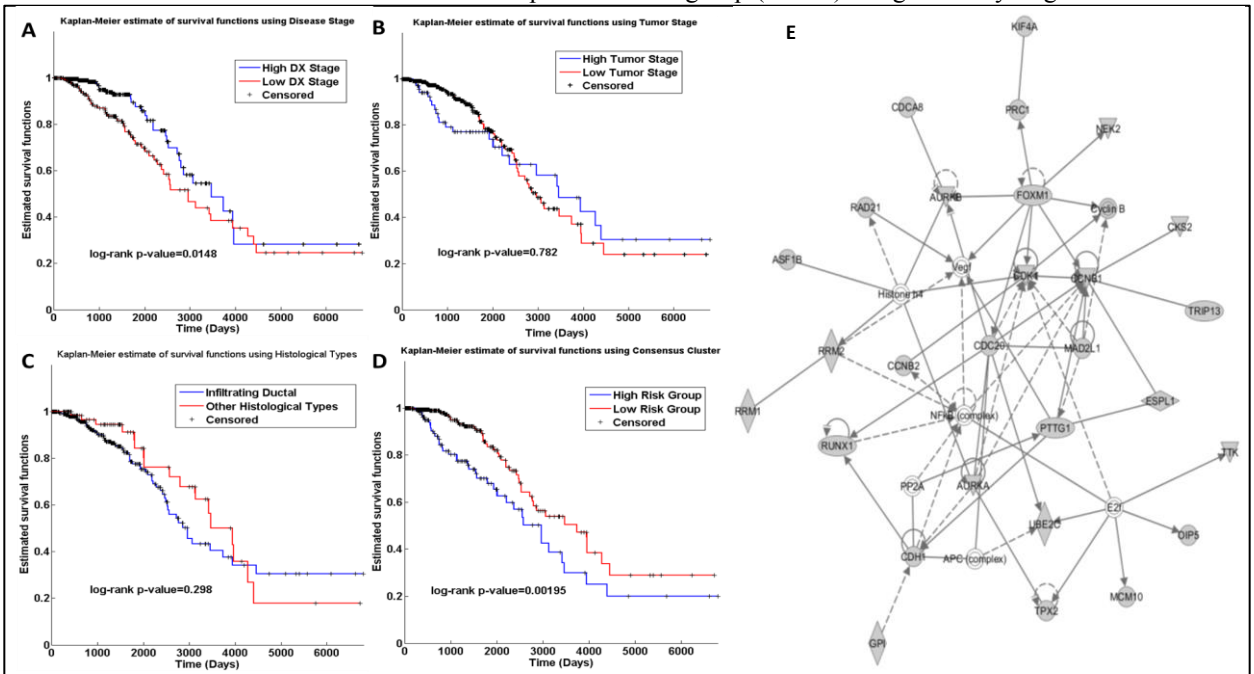


Figure 4. Kaplan–Meier plots for classification of the whole TCGA population. (A). Stratification based on patient disease stage. (B) Tumor grade. (C). Histological type. (D). Consensus clustering. (E). A network identified from the high-risk group.

rest (18.1%). Our observation confirmed the mutation of CDH1 and RUNX1 to be important in this subgroup of breast cancer, indicating that they could be the potential targets in this subtype.

4 Conclusion

We propose a new integrative patient classification method which aggregates both essential clinical information and multiple omic data. A breast cancer case study shows that miRNA expressions and mRNA expression can better utilize the associations between mRNAs and miRNAs, thus improves the classification on molecular level. The result of this case study also demonstrates the effectiveness of this consensus clustering patient stratification in breast cancer patient prognosis. In addition, the high-risk group of patients distinguishes from the rest in expression of 93 specific genes and mutation of two well-studied genes. The platform can be extended to any other types of cancer with any selection of biomarkers and clinical signatures. The extended data integration can provide hypothesis to identify both clinical and biological interesting cancer subgroups.

References

1. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 1996;6(7):639–645.

2. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 2011;12(2):87–98.

3. Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. *Biol. Direct*. 2012;7(1):21.

4. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747–52.

5. Brannon AR, Reddy A, Seiler M, et al. Molecular stratification of clear cell renal cell carcinoma by consensus clustering reveals distinct subtypes and survival patterns. *Genes Cancer*. 2010;1(2):152–163.

6. Cho S, Jeon J, Kim S Il. Personalized medicine in breast cancer: a systematic review. *J. Breast Cancer*. 2012;15(3):265–272.

7. The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70.

8. Strehl A, Ghosh J. Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* 2003;3:583–617.

9. Wang H, Shan H, Banerjee A. Bayesian cluster ensembles. *Stat. Anal. Data*. 2011:1–17.

10. Shen R, Mo Q, Schultz N, et al. Integrative subtype discovery in glioblastoma using iCluster. *PLoS One*. 2012;7(4):e35236.

11. Chaudhuri K, Kakade SM, Livescu K, Sridharan K. Multi-view clustering via canonical correlation analysis. *Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML'09*. 2009:1–8.

12. Lock EF, Hoadley K. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl.Stat*. 2013;7(1):523–542.

13. Saghafi B, Rajan D. Multi-view clustering of visual words using canonical correlation analysis for human action recognition. *2010 Ninth Int. Conf. Mach. Learn. Appl.* 2010:661–666.

14. Laura J, Veer LJ Van, Dai H, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(345).

15. Volinia S, Croce CM. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. *Proc Natl Acad Sci U S A*. 2013;110(18):7413–7.

Improving the Translation of Model Organism Research into Disease Diagnostics

Nicole L. Washington, Ph.D.[1], Melissa A. Haendel, Ph.D.[2], Sebastian Köhler Ph.D.[3], Suzanna E. Lewis, M.S.[1]; Peter Robinson, M.D.[3], Damian Smedley, Ph.D.[4], Christopher J. Mungall, Ph.D.[1]
[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]Oregon Health & Sciences University, Portland, OR; [3]Institut für Medizinische Genetik und Humangenetik, Charité - Universitätsmedizin Berlin, Berlin, Germany; [4]Wellcome Trust Sanger Institute, Hinxton, UK

Summary
In order to determine the underlying mechanism of a disease, animal models can often elucidate the biological underpinnings of the phenotype. We present our findings on the distribution, significance, and information characteristics necessary to enable translation of model organism research into disease diagnostic clinical applications using an ontological approach.

Background Determining the underlying mechanism of a novel and/or rare genetic disease, as well as identifying an appropriate model to study, are among the first steps toward discovery of a treatment and cure. Since such diseases do not have the power of large population sizes to perform genomic correlations, we must rely on other informational means to uncover the basis of their phenotypes. Toward this end, we can use the wealth of genotype-phenotype associations derived from model organisms and *in vitro* systems to elucidate genetic underpinnings. For example, "Parkinson's disease" can be broken down into a constituent set of phenotypes from the Human Phenotype Ontology (HPO), such as "Dysarthria", "Bradykinesia", "Dystonia", etc., and thereby used as a query set against ontologically encoded model system phenotypes. We have shown that comparison of human diseases with animal model phenotypes using the OWLsim algorithm (OWLsim.org) enables retrieval of existing curated model-disease relationships, as well as the orthologs of known disease genes[1,2]. This approach can be used alone or as a complement to whole-genome or exome analysis[3]. However, identification and prioritization of a relevant set of candidates for further study demands a sufficiently discriminatory set of phenotypes of the human disease in order to distinguish its given phenotypic profile from the rest of the corpus. Here, we investigate the necessary and sufficient information characteristics along several axes required to identify disease-gene relationships based on phenotypes alone.

Methods We integrated and analyzed semantically curated phenotypic characteristics and their properties of more than 7,000 genetic diseases from OMIM, Decipher, and Orphanet, together with the catalog of 47,000 mouse and 14,000 zebrafish genotypes with curated phenotypes from MGI and ZFIN, respectively. We used our derived cross-species phenotype ontology *uberpheno*[5] and Information Content (IC)-based measures of disease/model annotation specificity, to develop a formal statistical similarity measurement[1]. We used the upper levels of the HPO (e.g. divisions based on anatomical systems) as a classifier in order to build a statistical model, which was evaluated using iteratively less-specific synthetic patient profiles in order to identify the breadth and depth of annotations required to maximize the precision and recall of known genetic diseases and their animal models. Finally, we surveyed the annotation corpus to identify over and underrepresentation of domain knowledge, in order to better understand how the data could be used for cross-species analysis.

Results and Discussion We present our findings on the distribution and significance of annotations in our corpus with respect to attributes such as mono versus polygenic diseases, annotation source, and breadth and depth of annotations per high-level class. In addition, we present the performance of our model, and the annotation profile characteristics that are "informative" for identifying relevant animal models, similar diseases, and candidate genes when comparing phenotypes. For instance, we find that a greater number of annotations are required to return adequately unique models for nervous system diseases, because there are a disproportionately large number of nervous system phenotypes in human diseases and a relatively low level of graph complexity in this part of the HPO. The outcomes of this analysis are incorporated into an "annotation sufficiency" score available in the Monarch Initiative platform (www.monarchinitiative.org) as services that can be utilized by third party applications (such as Phenotips; phenotips.org) to inform users on how informative a set of annotations (e.g. diseases or models) are within the context of the whole corpus. This analysis will enable the development of minimum phenotype recommendations for clinical and model organism phenotype data publication and translation of model organism research into disease diagnostic clinical applications.

[1] Smedley *et al, Database*, May 9; 2013: bat025; [2] Washington *et al*, *PLoS Biology,* 7(11), 2009; [3] Robinson *et. al, Genome Res.* 2014 Jan 2; [4] Köhler *et al*, *F1000Research* 2013, **2**:30.

Evaluating CTSA Publication Output Using the Triangle of Biomedicine

Griffin M Weber, MD, PhD
Center of Biomedical Informatics; Harvard Medical School; Boston, MA, USA.

Summary: The NIH Clinical and Translational Science Awards (CTSA) program is developing a national infrastructure to support translational science. However, measuring its impact is challenging. This presentation uses a recently introduced bibliometric technique called the Triangle of Biomedicine to evaluate where publications citing CTSA grant numbers fall along the translational spectrum.

Introduction and Background: Translational science has been described in many different ways, including the frequently used qualitative T1-T4 classification; however, consensus on a precise definition has not yet been reached. The author recently introduced a bibliometric method called the Triangle of Biomedicine, which maps the 22 million biomedical journal articles in the National Library of Medicine's PubMed bibliographic database to a triangle, whose corners represent research related to animals (**A**), cells and molecules (**C**), and humans (**H**). The position of an article on the graph is based on its Medical Subject Heading (MeSH) topics, and translation is defined as movement of a collection of articles, or the articles that cite those articles, towards the human corner. This provides a way of determining the degree to which an individual scientist, organization, funding agency, or scientific field is producing results that have potential impact on human health, and calculating the amount of time it takes.

Methods: NIH ExPORTER was used to obtain more than 11,000 articles in PubMed citing CTSA grant numbers. We mapped these to the Triangle of Biomedicine to calculate a variety of metrics to measure their translational impact. More traditional bibliometric analysis methods were also used to evaluate the articles. Articles are grouped into disciples based on the journals they are in, using the approximately 100 Broad Journal Headings that NLM assigns to journals in PubMed.

Results: Individual CTSA institutions vary in their position within the Triangle of Biomedicine, with some having nearly 100% of their articles related solely to humans (**H**), and others being midway between humans and cells/molecules (**CH**). Few CTSA publications involve animals, which is surprising since animal studies are often a precursor to clinical trials. Combined, 95% of CTSA publications are about humans, 28% are about cells/molecules, and 13% are about animals. (The sum is greater than 100% since an article can belong to more than one category.) In contrast, articles from R01 grants are close to the center of the Triangle, with a larger percentage of basic science research. On average, CTSA publications have 35% more coauthors and are cited 54% more often than other articles published in the same journals in the same years. The most common discipline of CTSA publications is neurology. However, when compared to the distribution of disciplines in all of PubMed, metabolism and endocrinology are the most overly represented disciplines in CTSA publications.

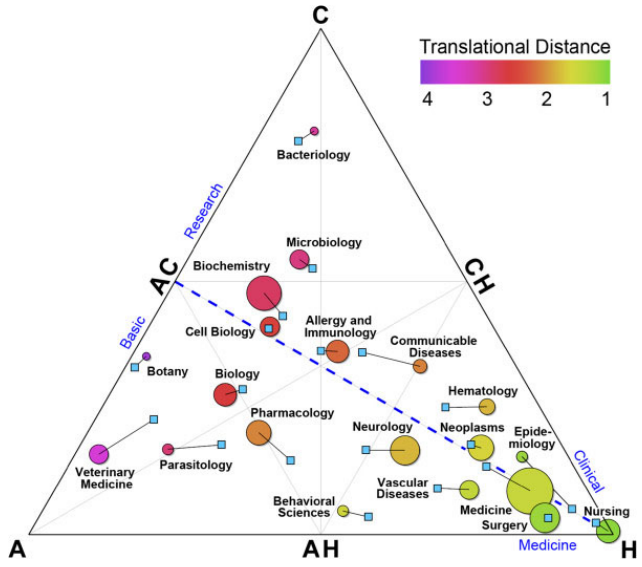


Figure 1. Triangle of Biomedicine. Circles represent different disciplines (not all are shown). Small squares are articles that cite those disciplines. Translational distance is the number of citation generations to get to a human study.

Discussion: The Triangle of Biomedicine provides a novel way to identify translational science and evaluate the impact of the publication output from the CTSA program. This study is limited by the fact that not all publications that benefit from CTSA cite the grant number and the number of citations listed in PubMed is considerably lower than in commercial citation databases. This may be addressed in the future with access to CTSA annual progress reports and data from Web of Science or Scopus.

References: Weber GM. Identifying translational science within the triangle of biomedicine. J Transl Med. 2013 May 24;11:126. PMCID:PMC3666890

PGxpress: a Pharmacogenomic Mobile Website

Michelle Whirl-Carrillo, PhD[1], Ryan Whaley, BS[1], Russ B. Altman, MD, PhD[1], Teri E. Klein, PhD[1]
[1]Stanford University, Stanford, CA

Introduction

PGxpress is a pharmacogenomic mobile website designed to provide users with a streamlined, adaptable interface to the pharmacogenomic knowledge in PharmGKB. PharmGKB (www.pharmgkb.org) is an online pharmacogenomics resource that contains manually curated information from the published literature, genotype-based drug dosing guidelines and annotated drug labels. The website component of PharmGKB is the primary entry point for accessing this knowledge and is designed to present complex information in an organized manner for both browsing and targeted searches. However, it can be challenging for users seeking the highest impact information to use the website outside of a traditional desktop web browser. PGxpress enables users to quickly access key PharmGKB knowledge with a lightweight, legible display on any size screen, including mobile devices like smart phones and tablets.

Methods

PGxpress is a modern web application based on HTML, Javascript, and CSS. It uses a traditional model-view-controller (MVC) pattern architecture and the PharmGKB REST API. The REST service on the server queries domain objects that are already in use on the PharmGKB site (e.g. drugs, dosing guidelines, clinical annotations) and transforms them into JSON format using the Jackson JSON processor before serving them to PGxpress. The PGxpress javascript controller layer is based on the AngularJS toolset. The view layer of PGxpress uses AngularJS directives and modules to bind data gathered in the controller to HTML elements on the page. It uses modernizr to support older rendering engines and the Twitter Bootstrap framework for visual layout, scaffolding, and UI element styling. Form elements such as select, textarea, etc. are all styled by the bootstrap styling rules.

Results

The PGxpress homepage has a simple, Google-type search box. Users can enter drugs, genes or variants to retrieve knowledge from PharmGKB. PGxpress retrieves any clinically relevant information and presents it in a condensed format at the top of the page and research-related variant annotations are accessible at the bottom. Available dosing guidelines with short summaries and drug label information from the FDA and European Medical Agency are listed. This page presents a quick overview of the level of evidence available for known pharmacogenetics associations and gives the user a quick sense as to the pharmacogenetic significance of a drug/gene before going into further detail. Following any of the provided links on the drug or gene page yields more in-depth information about the topic. A navigation bar at the top of each page allows quick maneuverability within and between pages. The pull-down menu displays all of the information types, with counts, for the object of interest. The user can move between dosing guidelines, labels and annotations. Additionally, a link to the full PharmGKB website is at the bottom of every page if the user is looking for more information about a topic.

Discussion

PGxpress provides a streamlined interface to the great volume of diverse pharmacogenomic information contained in the PharmGKB knowledge base. Simplified search functions and high-level summaries, together with a custom-formatted display, provide direct access to complex information on a mobile device. Researchers and clinicians live in a world where instant access to relevant information is a necessity, yet access to a computer is not always available. People routinely use smart phones and tablets in lieu of computer workstations and laptops. Hence the need for pharmacogenomic information in a format that is usable regardless of the screen size of the electronic device. In fact, mobile access is essential in the modern web environment regardless of the field of study, and PGxpress brings that capability to PharmGKB.

(NIGMS R24 GM61374)

Using a Biobank Linked to Electronic Medical Records to Identify Non-Specific Clinically Associated Genetic Variants

Laura K. Wiley, Robert Goodloe, Eric Farber-Eger, Dana C. Crawford, William S. Bush
Center for Human Genetics Research, Vanderbilt University Medical Center, Nashville, TN

Abstract

We hypothesize that leveraging information on patient populations to define more specific phenotypes or to evaluate comorbidities may improve our ability to detect functional genetic variants. This pilot study used principal component analysis of ICD-9 code histories and PheWAS to identify four non-specific clinically associated variants. The precise implications of the associations are unclear; however this approach may be useful for prioritizing and annotating variants with potential clinical impact.

Introduction

Traditional genetic epidemiology studies typically investigate only one or just a handful of phenotypes. Biobanks linked to electronic medical records (EMRs) have expanded the density of phenotype capture for a group of individuals. However the majority of studies that use these resources still rely on a case/control study design of carefully selected phenotypes. Genetic studies have identified numerous variants with small to moderate effect sizes, and are fast approaching a point where it is not feasible to collect large enough sample sizes to provide power to detect additional low-effect variants. In this work, we attempt to harness the power of the extensive phenotypes available in EMRs to identify variants with non-specific clinical effects. Genetic changes to some biological mechanisms may impact a spectrum of clinical traits on a low level, influencing co-morbidities, altering response to treatment, or affecting an underlying collection of medical problems that are not yet defined as "diseases". While the exact causal path of genetic effect on disease may be difficult to elucidate from this approach, the variants identified are of potential clinical interest and thus worthy of further investigation. In this pilot study we explore using the data reduction method of principal component analysis to identify variants with general clinical associations. We further investigate the effects of these variants on individual phenotypic outcomes using a phenome-wide association (PheWAS) approach.

Methods

We performed principal component analysis on a Boolean representation of the entire ICD-9 code history of 11,166 African American patients from EAGLE-BioVU, a cross-section of minority populations in the Vanderbilt DNA Biobank. We then regressed the top clinical principle component on each of 140,646 additively encoded single nucleotide polymorphisms (SNPs) passing quality control, adjusted for gender and the top three ancestry principle components. Significant SNPs passed a Bonferroni-corrected significance threshold of 3.6×10^{-7}. We then performed PheWAS on these variants using ICD-9 code groups using a significance cutoff of 2.6×10^{-5} as described in [1].

Results

Seven SNPs were associated with the first clinical principle component. Four of these, rs60661860, rs8036138, rs10949910, and rs12734338, had significant associations to 5, 7, 18, and 49 phenotypes respectively. Interestingly all variants were associated to cancer and cancer related phenotypes. There was also a cluster of phenotypes that appear to capture individuals with metabolic syndrome that were also associated to rs10949910 and rs12734338. Three SNPs (rs8036138, rs10949910, & rs12734338) alter a variety of regulatory motifs according to Haploreg v2. These SNPs were originally selected to tag copy number polymorphisms (CNPs) found in the Welcome Trust Case Control Consortium. Additionally, previous GWAS suggest further pleiotropic associations of rs12734338 to ALS ($p=8.9 \times 10^{-4}$) and HIV-1 viral load ($p=1.46 \times 10^{-5}$) [2].

Conclusion

Although the precise clinical implications of the identified variants are unclear, they do appear to influence phenotypes characterized by complex combinations of ICD-9 codes. These non-specific clinically associated variants may be useful for prioritizing and/or annotating variants in a variety of genetic studies.

References

1. Denny JC, Ritchie MD, Basford MA, et al. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010, May 1;26(9):1205-10.
2. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC Med Genet 2009;10:6.

CytoGPS (CytoGenetic Pattern Sleuth)

Zachary Abrams, BS; Arkaprava Pattanayak, BS; William Kenworthy; Lori Dalton, PhD; Philip R. O. Payne, PhD

Introduction and Background: Karyotyping tests detect chromosomal structural defects that can serve as genetic indicators of disorders or diseases. The International System for Human Cytogenetic Nomenclature (ISCN)[1] is a domain-specific language used to record these defects. ISCN encodes these chromosomal defects through visual microscope inspection. Currently, more karyotypic patient data exist, 65,000 in Mitelmen[2], than available genomic patient data. However, karyotypic data are difficult to analyze using existing computational methods due to their syntactic variability, information density and potential for human error.

Methods: To transform ISCN-encoded karyotypes to machine-readable constructs, several grammar-anchored parsers were generated programmatically using ANTLR (ANotherTool for Language Recognition)[3], an open-source software package. Context-Free Grammar (CFG), encoded using the Extended Backus-Naur form, was employed to write ISCN-specific production rules. These rules facilitated the tokenization of the ISCN karyotype and the transfer of tokenized elements into a biological model. This biological model represents the biological effect that a chromosomal aberration has on the individual, allowing clinicians and researchers to observe the biological effect rather than the aberration event. By combining location information and the type of event represented in the biological model, evaluating the impact of cytogenetic abnormalities at the more granular level of genes or gene products is possible. Employing the Ensembl[4] platform, we are able to retrieve all genes affected by a chromosomal aberration.

Results: The Mitelman database is the largest public repository of clinical karyotype data. Of the 65,000 karyotypes in Mitelman, our proposed system successfully parsed 95.5% of input karyotypes. Of the karyotypes parsed, 90% could be classified into the biological model. Using gene ontology enrichment and pathway analysis on the approximately 250 cancer groups outlined in the Mitelman database, we identified genes that were factors for their respective diseases.

Discussion and Conclusion: This proposed method allows the user to easily transition from karyotype to genetic-level data, facilitating a deeper level of clinical and research-relevant analysis as opposed to traditional approaches. These outcomes could lead to new discoveries in both treatment as well as diagnosis in a wide range of disease areas.

References:

1. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer (2013). Mitelman F, Johansson B and Mertens F (Eds.), http://cgap.nci.nih.gov/Chromosomes/Mitelman

2. Shaffer LG, Slovak ML, Campbell LJ. (2009). ISCN 2009: An International System for Human Cytogenetic Nomenclature (2009).

3. Parr, Terence (May 17, 2007). *The Definitive Antlr Reference: Building Domain-Specific Languages* (1st ed.), Pragmatic Bookshelf, p. 376, ISBN 0-9787392-5-6.

4. Ensembl, http://useast.ensembl.org/index.html

SPIRIT ML – A MACHINE LEARNING PLATFORM FOR TRANSLATIONAL RESEARCH

Srisairam Achuthan, PhD, Mike Chang, PhD, Ajay Shah, PhD, Joyce C. Niland, PhD
Research Informatics Division, Department of Information Sciences, City of Hope, CA

Abstract

A synergistic and flexible machine learning platform (SPIRIT ML) has been developed for analyzing biomedical datasets. An interactive interface, broad spectrum of learning models, different cross-validation methods and reporting metrics constitute the platform. SPIRIT ML was created with the goal of addressing varied data analysis problems in translational research.

Introduction

Machine learning methods have been applied to identify patient cohorts based on EMR data, identify malignant tumors based on image data, and for adverse drug surveillance based on publicly available databases, to name a few. Significant effort and time are spent when individual machine learning methods are applied to these kinds of problems via independent "one-off" deployments. We are building an integrative platform for research informatics called SPIRIT (Software Platform for Integrated Research Information and Transformation) that includes an interactive interface for applying Machine Learning (SPIRIT ML) methods. Based on a suite of algorithms, SPIRIT ML furnishes a streamlined approach to analyzing biomedical data. It serves as a comprehensive framework for clustering, classifying, and deciphering relationships among covariates encountered routinely in various biological and clinical settings.

Methods

SPIRIT ML provides unsupervised learning algorithms (hierarchical as well as non-hierarchical clustering), supervised learning algorithms (artificial neural networks, decision trees, support vector machines, random forests, logistic regression as well as naïve Bayes) as well as Bayesian network models using eight different algorithms to discover correlated features within a dataset. The raw data are automatically transformed via built-in options such as normalization and binning. The transformed data are distributed with a fixed percentage utilized for training, validating and testing purposes, based on user-defined values. Multiple cross-validation methods and well known reporting metrics have also been implemented. Open source R libraries and components from commercial applications such as MATLAB, Pipeline Pilot and Hugin form the backbone of the ML platform.

Results

We are beginning to apply SPIRIT ML to a wide range of research problems. Using Bayesian networks and supervised learning algorithms, a high school summer intern was able to determine the most important covariates that lead to tumor coverage by neural stem cells based on data derived from animal experiments. SPIRIT ML also is being used to identify patient cohorts based on clinical trial criteria utilizing de-identified patient characteristics available in our instance of the i2b2 open source software. With the aid of the ML platform potential risk factors among type II diabetics and certain cancer types (Liver, Pancreatic, Bladder, Breast, Ovarian, Colon, Uterus and Esophageal) are being investigated.

Conclusions

SPIRIT ML is a functional machine learning platform that can discover and reveal patterns in datasets. The underlying design of the platform is flexible enough to include machine learning models of choice, and facilitates comparison of results obtained by each model side by side. We intend SPIRIT ML to be an all inclusive platform so that machine learning methods developed in other open source packages such as WEKA can be incorporated with minimal effort via Web Services. We are able to assist multiple translational research projects that require data driven knowledge extraction.

Title: Towards the detection of riboswitch patterns in higher organisms using improved bioinformatic methods

Presenting and Corresponding Author: Danny Barash, Associate Professor, Department of Computer Science at Ben-Gurion University, Beer-Sheva 84105, Israel.
Co-author: Matan Drory, M.Sc. student, Dept. of Computer Science, Ben-Gurion University.

Summary: Riboswitches are RNA genetic control elements that provide a unique mechanism of gene regulation. They were originally discovered in bacteria where they are abundant. Only very few were discovered in eukaryotes. The discovery of more eukaryotic riboswitches, most ambitiously in humans, would enhance the implication for drug discovery. Towards this end, we are developing flexible methods that will improve riboswitch pattern detection.

Introduction and Background: Riboswitches are regulatory segments of an mRNA that provide a unique mechanism of gene regulation without the intervention of proteins. They were originally discovered in bacteria where they are abundant and only very few were discovered in eukaryotes. A riboswitch works by binding a small molecule, resulting in a change in production of proteins encoded by the mRNA. As a potential drug, it is possible to introduce fake metabolites to the cell that mimic the natural ones, find a riboswitch target, and turn off a gene. Because of the specificity of riboswitches to the genes they regulate and the specificity of the metabolite that triggers each riboswitch, it is possible to finely control the target of such a drug. Currently riboswitches are being tested as antibacterial drug targets, for example as a way to treat human bacterial infections. If more riboswitches could be discovered in higher organisms, and most ambitiously in humans, such findings would have important implications for drug discovery. Thus, the development of additional bioinformatic methods, which are specifically geared towards the detection of riboswitch patterns and can be applied in a variety of genomes, may lead to significant discoveries that are valuable to the medical field.

Methods: We are developing more flexible methods than the ones currently used, as well as novel structure-based methods, intending to improve riboswitch pattern detection. The flexibility is in the model, for example by the addition of dynamic programming for finding similarities of the query in the target instead of exact matching. Efficiency can be achieved by enhancement techniques such as memorization and the use of advanced data structures. The amount of sequence vs. structure similarity can be controlled. For example, a more heavily tilted structure-based method first identifies structural similarity and only then considers sequence similarity.

Results and Discussion: We were able to identify interesting riboswitch candidates in eukaryotes based on our developing structure-based method, such as a suspected purine riboswitch in Arabidopsis (Barash & Gabdank, RNA Biology, 2010). We are currently developing our methods and testing them on a variety of organisms. More ambitiously, we will apply these methods to human genomic sequence data for the purpose of riboswitch identification.

Metadata Approach to Leveraging Bio-Banking Information for Translational Research

Rimma Belenkaya[1], M.S., M.A., Vishwa Niranjan[1], M.S., Alexandre Peshansky[1], M.S., Xin Zheng[1], Ph.D, Parsa Mirhaji[1], M.D., Ph.D[1]

[1]Einstein-Montefiore Institute for Clinical and Translational Research, New York, NY.

Introduction

We introduce Einstein-Montefiore Bio-Repository Databank (EM-BRED), a metadata driven data integration and querying tool developed at Einstein-Montefiore Institute for Translational Research (EM-ICTR) for identifying bio-specimens for translational research projects based on phenotypic and clinical characteristics of their donor and pathological and bio-banking annotations of the specimen.

Background

Identifying biological specimens that qualify for a translational research project often represents a daunting task, largely due to the fact that specimen's donor and specimen annotations are stored in disconnected databases. Normally, the process starts from identifying patient cohort in electronic health records (EHR), then one by one, locating specimens in pathology or bio-bank databases and verifying that these specimens qualify for the research requirements. This process depends on availability of busy pathology personnel, is time consuming, and, as such, is quite limiting. A search engine to a repository that integrates donor and specimen annotations is a significant improvement to this process. Specimen and donor annotations vary between organs and diseases; so do specimen selection criteria for different studies that need them. Building a different indexing and search strategy for each project is inefficient. Our hypothesis is that an integration and retrieval strategy based on metadata representing disease and organ specific information enables a more generic and reusable approach.

Methods

We created a framework that consists of a metadata representation that supports a metadata-driven front-end and an integrated data repository.

The metadata represent components of user interface and their inter-relationships, define user interactions relevant to specimens and clinical domains, provide vocabularies to annotate clinical and phenotypic characteristics of specimens, and specimen collection processes, and enable integration of multi-source clinical and specimen data.

The metadata driven front-end dynamically generates user interfaces and navigations, implements query formulation based on user interactions, and represents search results, specimen details, and associated clinical annotations. The metadata determine appearance and behavior of the application, the user and the domain terminologies, and relevant search strategies specific to different diseases, specimen types, and organs.

The integrated data repository uses metadata to support representation and integration of data from a variety of sources: Excel spreadsheets, EHR, and Freezerworks bio-bank system. The data repository combines donor and specimen annotations stored in simple data structures described by the metadata. All data are stripped of PHI and restricted based on the current consent for specimen use.

Results

EM-BRED serves as a platform for specimen management at Liver Disease Research Center. Researchers select specimens using search criteria. The selection can be refined further using clinical and specimen annotations and available specimen quantity. EM-BRED also mediates communication of requests for available specimens, and specimen retrieval and release processes from the Einstein-Montefiore central bio-repository.

Discussion

EM-BRED is a big step forward in identifying and making specimens available for research. It has enabled EM-ICTR to capitalize on existing informatics and bio-repository resources to support translational research. The metadata is now being extended to support Lupus, Pathology and other departments.

However, its scalability is limited: UI and query complexity is restricted and data integration pipe lines are data source specific. These limitations are due to the current metadata structure.

We will address these limitations in the next versions of EM-BRED by adopting richer and more robust knowledge representation frameworks using Semantic Web.

The Los Angeles Data Resource (LADR): A Comprehensive Regional Data Resource for Research

Douglas S. Bell MD, PhD[1,2], Spencer SooHoo PhD [2,3], Robert Jenders MD, MS[2,4], Paul Fu, Jr., MD, MPH [2,5], Joshua Lee MD[6,7], Ayan Patel MS[1,2], Marianne Zachariah[1,2], Daniel R. Masys MD[8]
[1]UCLA Health, Los Angeles, CA, [2]UCLA Clinical and Translational Science Institute, Los Angeles, CA, [3]Cedars-Sinai Medical Center, Los Angeles, CA, [4]Charles Drew University, Los Angeles, CA, [5]Harbor-UCLA/LA Biomed, Torrance, CA, [6]Keck Medical Center of USC, Los Angeles, CA, [7]Southern California Clinical and Translational Science Institute, Los Angeles, CA, [8]University of Washington, Seattle, WA

The Los Angeles Data Resource (LADR) is a joint project of several Los Angeles health care provider organizations to use patient data in supporting collaborative clinical investigation and comparative effectiveness research. Our vision is to link data from a majority of health care providers in the Los Angeles region to provide a nearly complete picture of each patient's care and outcomes. To deliver value on the road to this vision, we are developing services in a stepwise fashion. The initial service is research cohort discovery, allowing investigators to conduct interactive queries for patients across the participating medical centers. The output of each cohort discovery query is a count of patients by site that match the query criteria.. Two additional services will then be created that enable different uses of the specified cohorts: cohort recruitment for prospective clinical research studies (LADR-CR) and the provision of cross-institutional limited data sets for retrospective comparative effectiveness studies (LADR-CE).

To formally establish the consortium of participating organizations, we have developed a Participation Agreement, a Policies & Procedures document and a user agreement. The Policies & Procedures specify that LADR is governed by an Executive Committee (EC) which consists of one voting member from each institution. Each institution may also include one to two non-voting members to ensure that persons with appropriate authority and expertise are represented in EC meetings. Changes to the Policies & Procedures require unanimous approval by the EC; the addition of new member institutions requires a two-thirds majority vote. Other decisions (e.g., types of data to extract, choices of data standards, and technical architecture) need majority approval by the EC. The EC also appoints and oversees working groups. Membership in LADR requires participation in the cohort discovery service. Member institutions may choose whether or not to participate in LADR-CR and/or LADR-CE. Investigators will have access to only those services that their institution participates in. For all services, we plan to keep member institutions in full control of their data by using federated approaches to run queries, conduct recruitment and assemble data sets for analysis. In addition, a Data Steward will audit and report to the EC on LADR use by all users, calling out any potentially inappropriate patterns.

To date, UCLA and CSMC have formally joined the Consortium and the initial SHRINE network for cohort discovery is expected to go live by March, 2014. Data on 6.5 million patients will be available from these initial participants, updated monthly. Data elements will include demographics, ICD-9 diagnoses and procedures, results for the 100 most-ordered lab tests, and medications (outpatient orders, inpatient administration), both mapped to RxNorm ingredients. In the next year, we expect to add Charles Drew University, the University of Southern California, the Los Angeles County Department of Health Services, and other community network members.

The LADR-CR system will harmonize patient recruitment methods among the Consortium members participating in particular studies. Each study's outreach will still require IRB approval and involvement from a responsible principal investigator at each participating organization. No member organization will directly recruit patients from another member's organization. LADR-CE will use "private record linkage" algorithms to join data for the same patient across different institutions without requiring any institution to release the patients' individual identifiers. By default, the site of care delivery will not be identified in LADR-CE data sets, even in coded form, unless an exception is approved by the Executive Committee, due to a compelling research need.

The LADR Consortium provides a model for joining electronic health record data from regional competitors to accelerate clinical and comparative effectiveness research. Health information exchanges (HIEs) also aggregate clinical data from organizations within a region, thus the use of HIE data could represent an alternative approach to providing regional data for research. However, the growth of many HIEs has been slowed by challenges of governance and trust among participants. The LADR governance model and architecture provide an example of success in fostering a high level of trust and collaboration among disparate, competing health care delivery organizations in the same region. The success of LADR will be evaluated based on its use in grant applications, accelerating clinical trial recruitment, and producing data for published comparative effectiveness studies.

How Bipartite Network Visualizations Complement Ingenuity Pathway Analysis: A Case Study in Methylation Related to Preterm Births

Suresh K. Bhavnani PhD[1], Bryant Dang BSc[1], Maria A. Caro MSc[1], Ram Menon PhD[2]
[1]Inst. for Translational Sciences, [2]Department of Obstetrics and Gynecology, UTMB, TX

Abstract

Although bipartite networks have been effective in identifying how significant biomarkers are associated to different subsets of patients, little is known about how domain experts use such patterns to infer biological pathways. Here we present a case study to elucidate how patterns from a bipartite network visualization were used in conjunction with Ingenuity Pathway Analysis (IPA) to infer pathways related to methylation in spontaneous preterm birth. The results suggest that because the network made explicit the inverse symmetrical relationship among cases/controls and methylation sites, the visualization helped to rapidly integrate pathways related to cases and to controls into a unified pathway hypothesis for spontaneous preterm. These results elucidate the complementary role that bipartite networks can play in inferring biological pathways from databases such as IPA.

Introduction

While bipartite networks have been effective in helping comprehend complex associations among subjects and biomarkers, little is known about how such information is used by domain experts to infer biological pathways. We therefore posed the question: *What information in a bipartite network of cases/controls and methylation sites is useful in identifying biological pathways from IPA?*

Method

In a recent study[1], we generated a bipartite network of 22 preterm cases (24-34 weeks gestational age) and 28 controls (>39 weeks gestational age) and the top-10 significant DNA methylation sites from a whole-genome study of cord blood of African American subjects. As shown in Figure 1, the network and subsequent cluster analysis revealed a patient cluster on the left with all but 2 cases that were hypermethylated (represented by dark gray edges) at 7 sites, and a patient cluster on the right with mostly controls that were hypermethylated at 3 sites. This network, along with IPA-generated pathways for each subset of methylated sites was provided to an expert in preterm biology. The expert was asked to think aloud while he attempted to identify the pathways, and his verbal protocol was recorded and analyzed to identify the steps taken to arrive at a hypothesis.

Results and Conclusion

The domain expert first attempted to analyze the IPA-identified pathways for 7 hypermethylated sites in the left cluster consisting of most cases. Unfortunately, none of the pathways appeared to be meaningful in preterm. Next, he analyzed the IPA-identified pathways for the 3 hypermetylated sites in the right patient cluster consisting of mainly controls. Here he determined that 2 hypermethylated sites (cg23754392, cg25592206) on genes BMI1 and CDKN2C respectively, which could be downregulated (due to being hypermethylated) resulting in the upregulation of TP53 (a known tumor suppressor identified by IPA) leading to normal cell senescence required for the normal rupture of the placenta during labor. Because these very sites were hypomethylated (represented explicitly by the light gray edges that connected the cases to these two sites) in most of the cases, he inferred that the opposite might hold for the cases: hypomethylation of the same 2 sites would lead to expression of BMI1 and CDKN2C, leading to the suppression of TP53 which in turn would result in minimal or absent cellular senescence, requiring surgical rupture of the placenta during preterm labor. Having determined a plausible role of cellular senescence in preterm, he reexamined the genes related to the left cluster of hypermethylated sites in the cluster of most cases. This led to a focus on BCL9 and IRF8, both of which are cell cycle promoters. He therefore unified the two insights by concluding that the absence of the pathway related to cellular senescence, in combination with the presence of a pathway that promoted cell cycle might be responsible for the preterm cases. These results elucidate how bipartite networks, which can explicitly represent inverse symmetrical relationships, can aid in rapidly integrating pathways related to cases and controls into a unified pathway hypothesis.

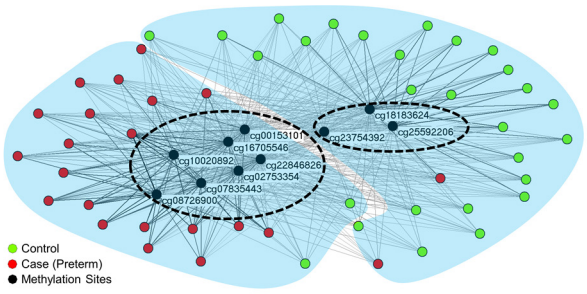


Figure 1. Bipartite network of 50 cases/controls, and top-10 significant methylation sites. Subject clusters are shaded in blue; methylation site clusters are marked with dotted ovals.

Acknowledgements. Funded by CTSA (UL1TR000071).

References

1. Bhavnani et al. Methylation Differences Reveal Heterogeneity in Spontaneous Preterm Birth Pathophysiology: A Visual Analytical Approach (*in press*).

Visualizing Clinically Similar Phenotypes

Charles Borromeo, M.S. 1, Nicole L Washington, Ph.D. 2, Christopher J Mungall, Ph.D. 2, Jeremy U. Espino, M.S. M.D. 1,Melissa Haendel, Ph.D. 3, Damian Smedley, Ph.D. 4, Jules Jacobsen, Ph.D. 4, Harry Hochheiser, Ph.D. 1
1 University of Pittsburgh, Pittsburgh, PA; 2 Lawrence Berkeley National Lab, Berkeley, CA; 3 Oregon Health & Science University, Portland, OR; 4 Wellcome Trust Sanger Institute, Hinxton, UK,

Summary: Numerous tools for exploring diagnoses rely on the ability to compare clinical phenotypes across patients. These inquiries can be further enhanced with comparative phenotypes from animal models. Here we present novel semantic visualization methods to aid clinical phenotyping through the incorporation of cross-species data.

Introduction: To support differential diagnosis through exploration of heritable traits and variant-disease hypotheses, it is necessary to compare phenotypes across patients. Comparison with animal models can further aid clinical phenotyping, gene-function testing, and disease treatment hypothesis testing. We have previously validated the results of these comparisons by comparing human phenotype annotation sets against mouse models. Semantic, ontology-based approaches using correspondences in phenotypes support detailed comparison across patients or species, but the results can be challenging to interpret. The Monarch Initiative (www.monarchinitiative.org) aims to provide interactive visualizations that support exploration of similarities between phenotype profiles for humans and corresponding model organisms.

Methods: Given a clinical phenotypic profile (e.g. a set of phenotypes from the Human Phenotype Ontology) that represent either a known disease or an undiagnosed disease, we use ontology-based similarity methods (OWLSim, www.olwsim.org) to identify similar model system characteristics (phenotypes), and specific models (mouse strains, cell lines, etc.) associated with the resulting profiles. Interpretation of these results involves the challenging integration of correspondences between sets of human and/or model phenotype descriptions encoded in multiple ontologies. We used the D3 and jQuery Javascript libraries to develop interactive graphical displays to aid in this interpretation. Established information visualization techniques including color-coding, mouse-over highlights of related information, and multiple coordinated views are used to display data and support interactive navigation. Ontological structures are used to support navigation and interpretation of inferred relationships between human phenotypes and candidate models.

Results: Our initial designs use parallel-set columnar views to show mappings between human and animal phenotypes, in combination with a grid-based display of correspondences between phenotypic descriptors and specific models. Coordinated highlighting of phenotypes-in-common use ontology relationships to illustrate how the correspondences were made. The parallel-set visualization provides a clear overview of how well a given set of model phenotypes recapitulates a human disease, while the grid-based display provides the ability to compare numerous models at once. The user can click between the two, providing detailed information about any given match as well as the reasons for matches across a set of highly similar models. The D3 and jQuery libraries support definition of these tools as widgets that can easily be adapted and reused in other websites.

Discussion: Interpretation of relationships between items in rich phenotypic ontologies raise interesting challenges in determining appropriate granularities for displays of similar items. “Rolling-up” multiple annotations, either via ontology structures or through additional structural calculations can provide greater clarity in display with minimal loss in fidelity. Our future research will also include comparative design evaluations and iterative development based on user feedback to improve the visualizations.

Population-Specific Manifestation of Insulin Signaling/Action Pathways: A Case Study of Chronic Metabolic Diseases in Colombians

Maria A. Caro MSc[1,2], Bryant Dang BS[1], Gabriel Bedoya MSc[2], Suresh K. Bhavnani PhD[1]
[1]Inst. for Translational Sciences, UTMB, USA; [2]GENMOL, Univ. of Antioquia, Medellín, Antioquia, Colombia

Abstract

Although the dysregulation of insulin signaling/action pathways are a well-known phenomenon in metabolic diseases such as type 2 diabetes and dyslipidemia, the manifestation of such pathways have a complex interaction with population-specific variables including life-style and ancestry. Here we use bipartite networks to analyze how SNPs on genes linked to insulin resistance, are associated with key demographic and clinical variables in Colombians with metabolic disease. The results revealed subsets of patients that were strongly associated with a heterogeneous subset of genes related to insulin resistance, with significantly high and low triglyceride levels explaining two of those associations. The results demonstrate how bipartite networks of genes from known pathways can rapidly elucidate their population-specific manifestations, with the goal of translation to contextually relevant therapeutics.

Introduction

While the role of insulin signaling/action pathways has been well-studied across chronic metabolic diseases (type 2 diabetes, obesity, hypertension, and dyslipidemia), their association with population-specific variables such as life-style and ancestry have yet to be fully elucidated. We therefore posed the question: *How do single nucleotide polymorphisms (SNPs) in candidate genes involved in insulin signaling/action, co-occur across Colombian patients with one or more metabolic diseases?*

Method

Colombian patients (n=340) with one or more metabolic diseases in the age range of 20-84 were genotyped for 10 SNPs on 6 candidate genes known to be associated with insulin signaling/action pathways[1]. Furthermore, we recorded clinical variables (triglycerides, waist circumference) and demographic variables (admixture, age, gender and socioeconomic status) for each patient. These data were analyzed using a bipartite network where nodes represented patients or SNPs, and edges represented the genetic association between each patient-SNP pair using the recessive model. We used bipartite modularity to determine the clusteredness (in comparison to random networks of the same size) of patient or node clusters, and the Mann Whitney *U* test to analyze which clinical and demographic variables were significant across the molecularly-defined patient clusters.

Results and Conclusion

As shown in Figure 1, the bipartite network was highly clustered (patient modularity= 0.382, SNP modularity=0.488), with each of the 5 patient clusters being associated to 1 or 2 SNPs. This result suggests that in the Colombian population, there are many entry points into the insulin signaling/action pathways. For example, Patient-Cluster-A was strongly associated with two SNPs on the KLF14 gene which regulates gene expression in adipocytes; Patient-Cluster-C (green nodes) strongly associated with two SNPs on the GCKR gene involved in the regulation of glucose phosphorylation in liver cells. Furthermore, Patient-Cluster-A (red nodes) had significantly lower triglycerides (U=6749.5, $p<0.01$, two-tailed) suggesting the presence of metabolic dysfunctions in absence of hypertriglyceridemia. In contrast, Patient-Cluster-C had significantly higher triglycerides compared to the rest (U=5872.5, $p<0.0005$, two-tailed) suggesting dysregulation of lipids metabolism in liver. The other 3 patient clusters had no variables that were significantly different compared to the rest of the patients, suggesting that in addition to having heterogeneous entry points into insulin dysregulation, these patients' responses were also heterogeneous[1]. The network therefore enabled us to examine a complex set of associations in a known pathway but in a new population. In future research we plan to analyze other pathways in chronic metabolic diseases using the same approach to elucidate the similarities and differences of pathways across populations.

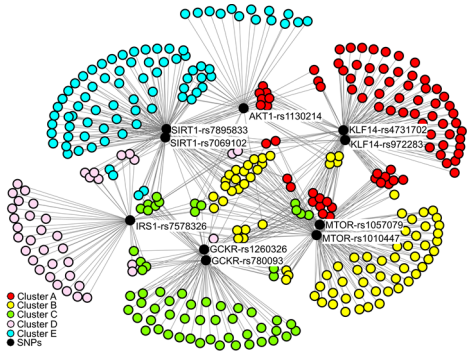


Figure 1. Bipartite network showing how candidate SNPs co-occur across Colombian patients with one or more metabolic diseases.

Acknowledgements. Funded by UTMB CTSA (UL1TR000071), and Colciencias 115-459-21587.

References

1. Blackett PR, Sanghera DK. Genetic determinants of cardiometabolic risk. *J Clin Lipidol* 2013; 7:65-81.

Predicting gene-level pathogenicity using variation in asymptomatic individuals

Christopher A. Cassa
Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

Corresponding Author: Christopher A. Cassa PhD, 77 Avenue Louis Pasteur, Boston, MA 02115, cassa@alum.mit.edu

Abstract

Whole genome sequencing (WGS) has the potential to improve medical care, but the methods to develop accurate clinical interpretations must be refined. The significance of novel genomic variation is generally characterized using *in silico* techniques that predict deleterious effects using evolutionary and functional considerations and the frequency of variation. We extend these techniques using a gene-based assessment of selective pressure using data from asymptomatic individuals in the Exome Sequencing Project (N=6,503). Using these features, we develop a gene-based Naïve Bayes classifier to assess pathogenicity of clinically important and benign variants. When trained variants (N=25,317) the classifier accurately predicts variant pathogenicity (AUC=0.798) and can predicts well in the separately ascertained HumVar dataset (AUC=0.761). These findings suggest that asymptomatic variation can be leveraged to predict variant pathogenicity.

Introduction

Whole genome sequencing (WGS) has the potential to improve medical care, but the methods to fully translate sequence data into accurate clinical interpretations remain to be defined [1]. The assessment of variants of unknown significance (VUS) poses a major challenge, as many of these variants are rare and have limited supporting phenotypic data to assist in disease classification. Even completely healthy individuals often carry variants that existing *in silico* predictive techniques suggest are pathogenic and that have been previously associated with disease [2].

Without computational predictive techniques to review variants in patients lacking clinical suspicion, it is difficult to determine whether these are clinically relevant findings, or false indications that may frighten patients and cause needless diagnostic workups and costly screenings [3]. This study focuses on the prioritization and pathogenicity assessment of newly observed variants. This is one of the most urgent needs in clinical genomic interpretation [4] as clinical labs [5,6,7] and direct-to-consumer groups [8] are already providing WGS interpretation.

We extend existing pathogenicity classifiers [9,10] using population data to measure signals of selective effect and protein fragility, by gene. We infer the tolerated variation in each gene in asymptomatic individuals to predict the sensitivity of each gene to new variation. If a gene tolerates many nonsense or missense variants in asymptomatic individuals, we would expect that novel nonsense or missense variants in that gene are less likely to be pathogenic. Together, the expected numbers of variants in each functional class, in each gene can serve as a proxy for selective effect.

Here, we develop a gene-based assessment for tolerance to different types of genomic variation. Using data from asymptomatic individuals (N=6503) in the Exome Sequencing Project (ESP) [11], we generate predictive features using variants of different functional classes and types in each gene. These data are then used to rank genes by expected number of variants, for different functional classes and minor allele frequencies, across two populations. We then use this data to train and validate a Naïve Bayes pathogenicity classifier.

Methods

Expected asymptomatic gene variation dataset

Using data from ESP, we calculate the number of each variant type that in each individual, by gene. Specifically, we calculate the number of heterozygous variants, homozygous non-reference variants, and compound heterozygotes. For each major type of variant described above, we stratify by population (European American, African American, or All). We also calculate these expected values for the functional classes described in ESP, e.g. missense or nonsense variants. We restrict our analysis to rare variants, with minor allele frequency (MAF) limits, ranging from 0.1 to 2.9%.

For each combination of the above characteristics (as described in Supplementary Materials Table 4,) we calculate the expected number of variants in each gene, in each individual, and generate a ranked list for all of the genes in that category. This full set of 174 features was then reduced to the 46 most informative features, which are used in this study, described in Table 1.

Table 1: Features used to classify variants. We describe the 46 major features per gene that are derived from the variation in the Exome Sequencing Project (ESP). For each functional class, variant-type, and population combination, we calculate the number of variants in each gene in asymptomatic individuals and also generate a gene rank-order list for that combination. We also include two additional features: the UniProt canonical transcript length and a de-novo mutation model.

Category	*Each combination of the following three categories:*				
Functional class	Missense	Nonsense	Synonymous	Mis./Synon.	Non./Synon.
Variant type	Exp. Heterozygotes by genotype	Exp. Homozygous by genotype	Exp. Homozygous by MAF		
Population	European American	African American	All		
Other	Protein length (canonical transcript length), relative replication time, and fraction of common variants				

For heterozygous variants, we use genotype count data provided by ESP to calculate the number of these variants. For homozygous alternative variants, we calculate two different values for each gene: the number of homozygous alternative variants derived from genotype count data, and the number of homozygous alternative variants derived from minor allele frequencies. This is the sum of the squares of the minor allele frequencies for each variant in each gene. For compound heterozygotes, we calculate the sum of all pairwise minor allele frequencies ($q_i * q_j$), excluding the case where $i=j$, for all variants within each gene.

Additionally, three other features are included: the length of the canonical gene transcript, the relative replication time of each gene, using smoothed data median of replication data (Koren et al 2012), and the fraction of common variants in each gene, as a signal of selective pressure on each gene. The fraction of all variants in each gene that has MAF > 0.01 and MAF > 0.05 is measured for each population group (AA, EA, and All).

Training and validation datasets

To develop and validate our classifier, we used three variant datasets, described in Table 2. The first is a custom set of variants derived from common ESP variants (benign) and uncommon HGMD variants (pathogenic). The second is the complete set of HumVar variants that can be mapped to a gene by UniProt accession numbers. The final dataset is designed for cross prediction with the custom ESP/HGMD dataset, and excludes variants from HumVar included in the custom dataset.

Table 2: Training and validation datasets.

	ESP and HGMD	**HumVar**	**HumVar excluding custom**
Benign class	ESP variants with PolyPhen2 = Benign and MAF > 0.03 (N=10,150)	HumVar benign dataset where UniProt canonical gene length available (N=18,498)	HumVar Benign not in custom dataset where UniProt canonical gene length available (N=11,874)
Pathogenic class	HGMD "DM" variants published after 2003 that are not observed in ESP (N=15,166)	HumVar pathogenic dataset where UniProt canonical gene length available (N=21,191)	HumVar pathogenic not in custom dataset where UniProt canonical gene length available (N=17,148)
Total	25,316 variants	39,689 variants	29,022 variants

Classifier validation and testing

We then use these data to develop an unsupervised classifier to predict variant pathogenicity. We train a Naïve Bayes classifier that uses the relative frequency of each feature to develop priors, with a LOESS (LOcally Estimated Scatterplot

Smoothing) window size of 0.5 using 100 sample points. For validation of predictive power, we perform a logistic regression and generate a shallow classification tree with a maximum node depth of 5 for each dataset.

Results

Classification tree

To demonstrate the predictive value of these features, we generate a simplified classification tree which classifies individual genes as either benign or pathogenic. Using the dataset derived from ESP and HGMD, we include only two background variation features: the rank of nonsense heterozygotes and missense heterozygotes. With only six leaves (distinct classes containing variants) and a maximum depth of 3 (the number of decisions that must be made before reaching a specific class), this tree has an AUC of 0.705. This demonstrates considerable predictive value at the gene level using just two background variant components.

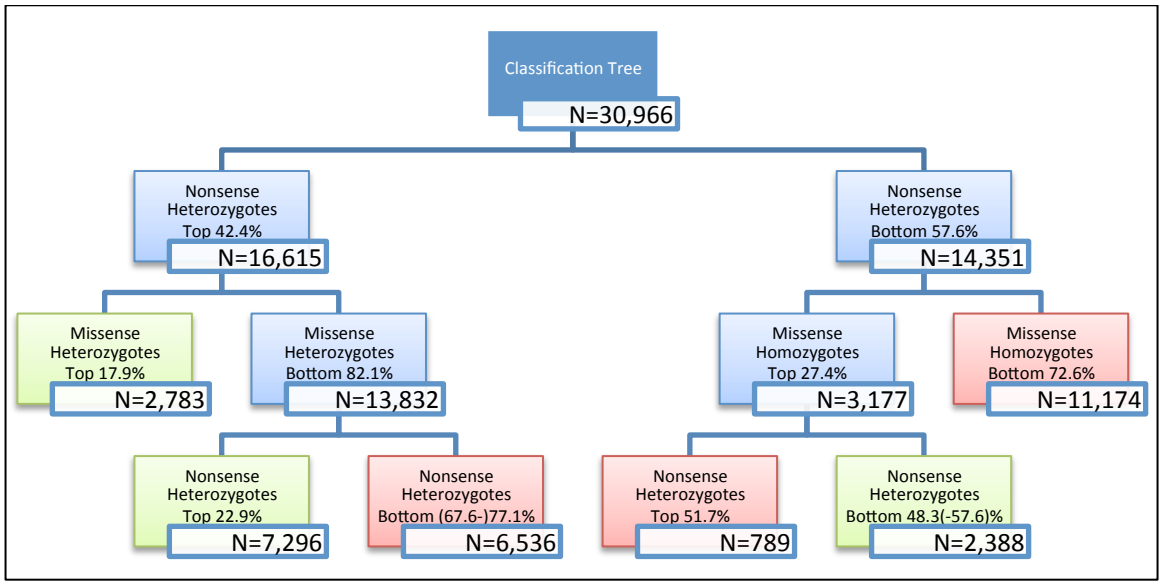


Figure 1: A simplified classification tree using only two features: the rank of nonsense heteryzygotes and missense heterozygotes

In this tree, we observe that genes with more nonsense heterozygous variants (top 42.4% of genes) are likely to be benign as long as they also have many missense heterozygous variants (top 17.9% of genes) or also fall within the top 22.9% of genes with nonsense variants. Conversely, genes with fewer nonsense heterozygous variants (bottom 57.6%) are likely to be pathogenic unless they have many missense variants (top 27.4%) and moderately few nonsense variants (bottom 48.3-57.6%). This implies that variants are more likely to be benign when they are observed in genes that accumulate and tolerate greater numbers of nonsense and missense variants, in general.

This performance on these core features is confirmed using a Naïve Bayes classifier that is trained on the ESP/HGMD dataset, which achieves an AUC 0.702 with only three features: the rank of nonsense heterozygotes, rank of missense heterozygotes, and rank of synonymous heterozygotes per gene.

Naïve Bayes classifiers and validation

We trained Naïve Bayes classifiers using the 46 most informative features with the custom dataset, the complete HumVar dataset, and HumVar excluding variants in the custom dataset. We confirm these results using a classification tree and logistic regression using the same features in each dataset. Each of these approaches demonstrates predictive value of the included features, described in Table 3.

Table 3: Results from training and validation of Naive Bayes classifiers

Method	ESP and HGMD	HumVar	HumVar excluding custom
Naïve Bayes	AUC: 0.7983	AUC: 0.7892	AUC: 0.7816
Classification Tree	AUC: 0.8050	AUC: 0.7994	AUC: 0.7936
Logistic Regression	AUC: 0.7886	AUC: 0.7728	AUC: 0.7675

Cross prediction with HumVar

The Naïve Bayes classifier that was trained on the custom ESP/HGMD dataset to predict variants was used to predict pathogenicity in the separately ascertained HumVar dataset (which excludes variants from the custom dataset.) Predictions made with the classifier that was trained on the custom set and tested on the HumVar set achieve an AUC of 0.7609. Predictions made using a classifier that is generated using the HumVar dataset and tested on the custom dataset are made with an AUC of 0.7817.

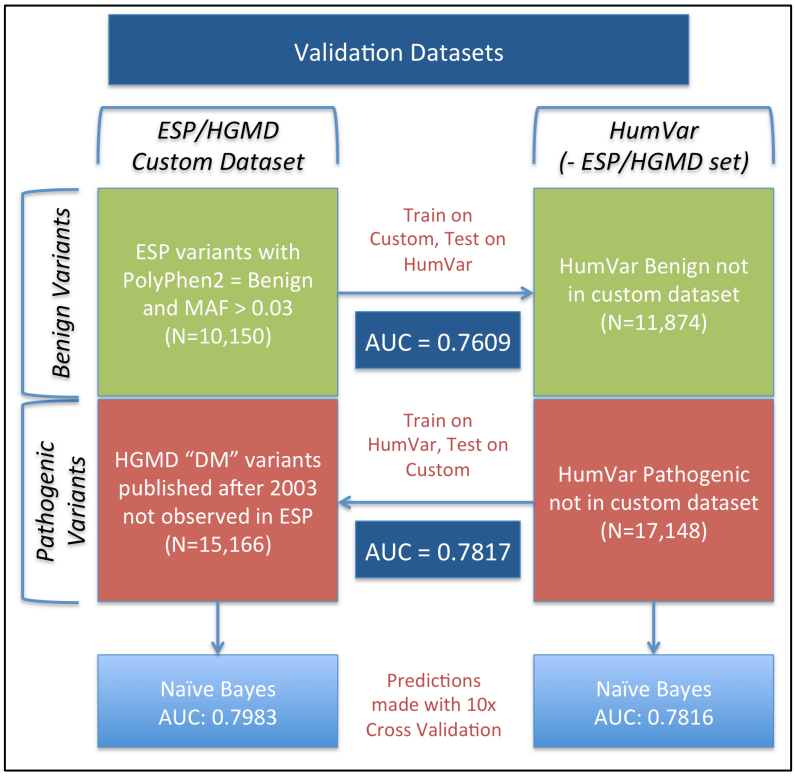


Figure 2: Cross validation between ESP/HGMD custom dataset and HumVar dataset that excludes any custom variants.

Discussion

Current symptom-driven genetic testing generally analyzes variants identified in a small set of disease-associated genes. Because these genes have been thoroughly studied before clinical testing, many variants that are identified can be quickly filtered out as known benign variation, and only a handful of variants are left for clinical consideration. This approach is not scalable to evaluate the 3 million variants found in the average human genome, and other strategies must be taken to infer and classify the pathogenicity of these variants.

Clinical interpretation of WGS data have already identified and attempted to mitigate disease risk [12] and demonstrate the urgent need for this filtering and prioritization research[13]. Many groups, including the ACMG, recognize that it is appropriate to share some findings in specific genes that are associated with disease [14], but this morbidity-based approach includes a very limited set of genes, without a complete list of variants. There are already an estimated 12,000 variants, genome-wide, that have sufficient clinical relevance and scientific validity for investigators to share them with research participants [15,16]. This means that there are currently over 12,000 variants that are appropriate to review and report. At present, there is no reliable, quantitative method to assess which of these variants is clinically significant or scientifically valid.

These expected numbers of variants may also be used to contextualize the interpretation of variants. For example, it may be helpful to know the expected frequency of a specific type of variant that has been observed, to help interpret whether it is likely that the variant of unknown significance is unusual. There are also potential clinical genetic diagnostics applications that use these predictive features. These features may improve existing analysis in morbid gene detection and gene-based burden testing, where likelihood ratios are not generally normalized using expected background variation. Similarly, these tables can help in mutational mapping by filtering noise, which reduces statistical significance in signals, and by amplifying signals that might otherwise go undetected. In the diagnosis of monogenic disorders, these techniques can help contextualize the importance of a single observation or prioritize a list of candidate variants. With a mutational mapping or burden gene-based approach, these data may also help identify associations between complex traits.

ESP provides whole exome sequences at high of coverage, however even with this large number of sequences, we do not have power to detect very rare variation. Future work may leverage a larger set of exomes for asymptomatic individuals, and larger pools of genetic and phenotypic data from participants in large longitudinal clinical studies [17,18,19,20,21,22].

Conclusion

This dataset and classifier will allow for expanded use of a large knowledge base of genetic variants in the clinical application of WGS findings. We provide data that may be used to filter results that are unlikely to be clinically impactful, protecting patients from invasive and unnecessary interventions and screening. This will also help prioritize the variants that are professionally reviewed by clinical geneticists, so time is spent reviewing the most promising candidate variants.

References

1. Brunham LR, Hayden MR (2012) Medicine. Whole-genome sequencing: the new standard of care? Science 336: 1112-1113.
2. Cassa CA, Tong MY, Jordan DM (2013) Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. Hum Mutat.
3. Kohane I, Masys D, Altman R (2006) The incidentalome: A threat to genomic medicine. JAMA 296: 212-215.
4. Green ED, Guyer MS (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470: 204-213.
5. genomeweb (2011) Baylor Whole Genome Laboratory Launches Clinical Exome Sequencing Test.
6. genomeweb (2011) Partners HealthCare Center's LMM to Introduce Clinical Whole-Genome Sequencing Interpretation Service in 2012.
7. Review T (2011) Making Genome Sequencing Part of Clinical Care.
8. Knome, Inc. Know thyself. Personal Human Genome Sequencing.
9. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, et al. (2001) Prediction of deleterious human alleles. Hum Mol Genet 10: 591-597.
10. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4: 1073-1081.
11. NHLBI (2012) NHLBI GO Exome Sequencing Project.
12. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. (2008) Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet 4: e1000167.
13. Biesecker LG (2010) Exome sequencing makes medical genomics a reality. Nat Genet 42: 13-14.
14. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, et al. (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15: 565-574.
15. Cassa CA, Savage SK, Taylor PL, Green RC, McGuire AL, et al. (2012) Disclosing pathogenic genetic variants to research participants: Quantifying an emerging ethical responsibility. Genome Res.
16. Fabsitz RR, McGuire A, Sharp RR, Puggal M, Beskow LM, et al. (2010) Ethical and practical guidelines for reporting genetic research results to study participants: updated guidelines from a National Heart, Lung, and Blood Institute working group. Circ Cardiovasc Genet 3: 574-580.
17. Benjamin EJ, Dupuis J, Larson MG, Lunetta KL, Booth SL, et al. (2007) Genome-wide association with select biomarker traits in the Framingham Heart Study. BMC Med Genet 8 Suppl 1: S11.
18. Genome-Wide Association Studies. In: genome.gov, editor.
19. Morton NE (2008) Into the post-HapMap era. Adv Genet 60: 727-742.

20. Cappuccio FP, Oakeshott P, Strazzullo P, Kerry SM (2002) Application of Framingham risk estimates to ethnic minorities in United Kingdom and implications for primary prevention of heart disease in general practice: cross sectional population based study. Bmj 325: 1271.
21. Colditz GA, Coakley E (1997) Weight, weight gain, activity, and major illnesses: the Nurses' Health Study. Int J Sports Med 18 Suppl 3: S162-170.
22. Empana JP, Ducimetiere P, Arveiler D, Ferrieres J, Evans A, et al. (2003) Are the Framingham and PROCAM coronary heart disease risk functions applicable to different European populations? The PRIME Study. Eur Heart J 24: 1903-1911.
23. NCBI (2012) database of Genotypes and Phenotypes (dbGaP).
24. EBI (2012) The European Genome-phenome Archive.
25. NCBI (2012) ClinVar.

Supplementary information

Table 4: Full list of features described in ESP background variation dataset. We generated multiple versions of the gene-based expected variation datasets using data from the Exome Sequencing Project (ESP). For each class, variant-type, MAF bin, and population combination, we calculated the expected variation in each gene in asymptomatic individuals and also generated a gene rank-order list for that combination.

Variant type	**Heterozygous**			**Homozygous alternative**			**Compound heterozygotes**	
Population	All			European American			African American	
Functional class	All	Splicing	Stop-gain	Stop-lost	Missense	Not Mod 3	Synon.	Other
MAF limit	0.1	0.5	0.9	1.3	1.7	2.1	2.5	2.9

Systematic Evaluation of the Applicability of Gene Expression-Based Therapeutic Target Validation

Bin Chen* (PhD), Hua Fan-Minogue* (PhD, MD), Weronika Sikora-Wohlfeld (PhD), Atul J Butte (PhD, MD)
Division of System Medicine, Department of Pediatrics,
Stanford University School of Medicine, Stanford, CA, 94305
(*: contribute equally)

Summary:
In this study, we systematically examined the general applicability of using differentially expressed (DE) genes as candidates for target validation. We found that the applicability of using DE genes as targets for drug development is disease-dependent and the majority of the DE genes are actually not known therapeutic targets.

Background:
Gene expression profiling is one of the most abundant and widely used data resources for understanding the molecular mechanism of disease. Identifying highly differentially expressed genes between normal and disease samples has been widely adopted to select therapeutic targets. However, whether DE genes are actually valid therapeutic targets for every disease has not been investigated systematically. Obtaining a systematic view of which disease has known therapeutic targets differentially expressed and which does not have could help determine the applicability of using DE genes for target validation for a certain disease.

Methods:
We retrospectively examined the expression of known disease therapeutic targets in the disease gene expression profiles. Known disease therapeutic targets were collected from Therapeutic Targets Database (TTD) and disease-related gene expression profiles were collected from Gene Expression Omnibus (GEO) and Array Express (AE). Disease concept IDs from the Unified Medical Language System were used to represent diseases. DE genes were computed using the methods RankProd and SAM respectively (FDR < 0.05, fold change > 1.5) and those genes were ranked based on fold change. The relationships between targets and DE genes were systematically analyzed and were further validated by external datasets including TCGA.

Results:
We collected 1001 disease-therapeutic-target pairs comprising 93 diseases and 611 targets. These diseases were profiled across 220 microarray datasets. 92.3% diseases have at least one DE gene. Among these, 43% diseases have at least one known therapeutic target differentially expressed. Restricting to the top 10% DE genes yields the worst recall, and selecting all DE genes is slightly better than other thresholds. The targets themselves had no significant preference to be in either up-regulated or down-regulated gene lists.

Discussion:
Our analysis provides a quantitative validation of current usage of gene expression in target validation, as well as guidance in target selection (e.g., direction, threshold). Solely using gene expression data for target validation may be not applicable for a number of diseases. Other types of data such as mutation, variation, gene copy number variation, protein-protein interaction and pathway should be employed as well.

Whole-Genome Sequencing Analysis Challenges and Solutions

Brian Conkright, MS, Krithika Bhuvaneshwar, MS, Michael Harris, MA,
Lei Song, MS, Yuriy Gusev, PhD, Subha Madhavan, PhD
Innovation Center for Biomedical Informatics
Georgetown University Medical Center, Washington, DC

Summary

Next-generation sequencing is a significant advancement that has opened new avenues for biomedical research. However, there are challenges in dealing with the size and complexity of whole-genome sequencing data, especially in a multi-sample setting. We explore solutions to these challenges using infrastructure and software that are readily available to most researchers.

Introduction and Background

As the cost of whole-genome sequencing (WGS) drops, disease studies that use hundreds or thousands of samples are becoming economically feasible. As compared to previous array-based methods, there is greater potential to explore the full complexity of the human genome—densely genotyped single-nucleotide polymorphisms (SNPs), copy number variations, and structural variations all available from a single, comprehensive data source. However, there are unique challenges in dealing with this data, which include ensuring quality throughout the analytical pipeline, managing numerous file formats, and effectively scaling computational resources to massive datasets. Additionally, some analysis challenges shared with previous generation platforms include: accounting for population heterogeneity, minimizing spurious associations, and mitigating batch effects. We have identified methods to address these challenges using cloud computing, publicly available software, and analytical best practices.

Methods

A striking feature of WGS datasets is the sheer size of files involved. To handle these massive datasets, we have made use of Amazon Web Services (AWS), in particular Elastic Cloud Compute (EC2) and Simple Storage Service (S3). EC2 instances are secured via public-key cryptography, while S3 is password protected. Additional security measures may be taken to produce a HIPAA compliant environment, such as event logging or file encryption.

Within this compute environment, we use software tools including Sickle, Bowtie, SAMtools, and GATK for quality control, read mapping, and variant calling of WGS data. Once variant call format (VCF) files are obtained, other downstream quality checks may include setting low-quality calls to missing, removing low-quality samples or variants, and checking for Mendelian inheritance errors, sex concordance, and significant departures from Hardy-Weinberg equilibrium.

In a multi-sample setting, tools such as PLINK and R may be scaled to run association tests on many cores within the compute environment using GNU parallel or other process-management software. Relevant tests include burden, chi-square, and logistic regression. Population structure may be accounted for using the top principal components in logistic regression, and similarly, batch effects may also be mitigated by including batch designation as a covariate.

Results and Discussion

We have found that these methods allow us to effectively manage and analyze multi-sample WGS data. For example, our WGS pipeline implemented in the Amazon cloud can do alignment and variant calling for a single sample in approximately 12 hours, and this solution scales to run additional samples in parallel. The standard data storage costs applied for AWS S3 storage. The current AWS costs can be obtained using the AWS calculator at http://calculator.s3.amazonaws.com/calc5.html. Given that each sample is run on a dedicated instance, the cost per sample is about $16. Further cost reductions are possible through the use of spot instances, which charge a fluctuating spot rate that is generally less than the fixed on-demand rate. Parallelization as described above reduced the runtime of association testing of millions of SNPs by an order of magnitude.

The demand for this type of processing and analysis will only grow as the cost of next-generation sequencing continues to decline, and more genomes become privately and publically available. Cloud computing is a sensible solution for researchers with variable budgets and computational demands. These computational and analytical practices can contribute to the production of timely, accurate results when working with WGS data.

Revealing Heterogeneity in Gene Regulation through Network Edge Coloring: A Case Study in Pediatric Pulmonary Infections

Bryant Dang BS[1], Shyam Visweswaran MD PhD[2], Asuncion Mejias, MD PhD[3], Rohit Divekar MBBS PhD[4], Suresh K. Bhavnani PhD[1]

[1]Inst. for Translational Sciences, UTMB; [2]Dept. of BMI, Univ. of Pittsburgh, Pittsburgh, PA; [3]Div. of Pediatric Infectious Diseases, Ohio State University, Columbus, OH; [4]Div. of Allergic Diseases, Mayo Clinic, Rochester, MN

Abstract

Although bipartite network visualizations have been effective in revealing heterogeneity in diseases (e.g., through patient node clusters representing subphenotypes), the resulting layouts often have many intersecting edges which can conceal important patterns such as gene regulation. Here we demonstrate the utility of coloring edges in a bipartite network to represent fold change (with respect to the controls), to help reveal differences in the gene regulation patterns among subphenotypes. The results suggest that colored edges representing fold change can help domain experts detect complex regulation patterns in bipartite networks compared to just expression values.

Introduction

Bipartite networks (which can simultaneously represent both patients and their molecular information such as gene expression), have helped to reveal heterogeneity in disease. For example, we used a bipartite network where nodes represented children less than 2 years of age infected with either influenza or respiratory syncytial virus (RSV), matched controls, 18 genes that were significantly expressed in both types of infection, and edges that represented normalized gene expression[1]. The network (Figure 1) revealed 3 clusters of patients: *core cases* that had high gene expression of 14 genes at the top of the network suggesting hyper-responsiveness, *periphery cases* that had medium expression of all 18 genes suggesting medium responsiveness, and 4 *control-like* cases that had a gene expression signature that was similar to the controls at the bottom of the network suggesting normal responsiveness. However, the network was too dense to comprehend overall patterns of gene regulation. We therefore posed the question: *Can colored edges in a bipartite network help to reveal differences in gene regulation across subphenotypes.*

Method

We median-centered all gene expression values (by dividing the median gene expression of the controls across all samples, including the controls, for each gene), and mapped the values to colors using the following range: <=0.2 (green), 1 (black), >=4 (red), where values above and below 1 represent up and down regulation respectively. As shown by the edges of the network in Figure 1, this range was used to color the edges to represent fold change. The network was presented to a domain expert in immunology to detect and interpret novel patterns related to gene regulation, and the Mann Whitney *U* test was used to quantitatively verify the significance of the pattern.

Results and Conclusion

The domain expert inspected the network and identified a new pattern related to gene regulation that was not salient in the network without the colored edges that was previously analyzed. As shown in Figure 1, the *core cases* (blue triangles and diamonds) strongly upregulated (mostly red edges) the 14 genes at the top, and strongly downregulated (mostly green edges) the 4 genes at the bottom. In contrast, the *control-like cases* had medium regulation (mostly dark edges) of both sets of genes. Statistically, the median expression of the core cases (Median=4) across the 14 genes on the top was significantly (U=625, p<.001) higher compared to the median expression (Median=0.26) of the 4 genes at the bottom. In contrast, the median expression (Median=1.54) of the control-like cases across the 14 genes on the top was not significantly different (U=15, p<.06) compared to the median expression (Median=0.93) of the 4 genes at the bottom. The results therefore revealed an important difference in the amplitude of genomic perturbation across subphenotypes providing a testable translational application to clinical outcomes. Coloring edges in a bipartite network therefore appear to be an effective way to reveal gene regulation differences amongst subphenotypes.

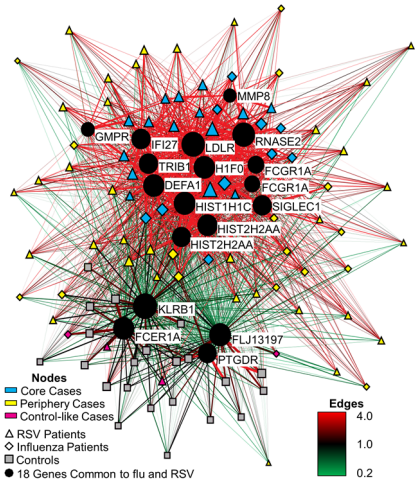


Figure 1. A bipartite network of patients with flu or RSV, 18 common genes, with edge length representing normalized gene expression, and edge color as fold change.

Acknowledgements: Supported by IHII, & NIH UL1TR000071 UTMB CTSA.

References

1. Bhavnani S.K., et al. Heterogeneity within and across Pediatric Pulmonary Infections: From Bipartite Networks to At-Risk Subphenotypes. *AMIA Summit on Translational Bioinformatics* (in review).

Whole Genome SNPs Data Analysis using parallel computing in the cloud

Andrea Demartini, MS[1], Davide Capozzi, PhD[1,2], Alberto Malovini, PhD[1,2,3], Annibale Puca, MD, PhD[4,5], Riccardo Bellazzi, PhD[1,2,3]

[1]University of Pavia, Pavia, Italy; [2]Biomeris s.r.l. , Pavia, Italy; [3]IRCCS Fondazione Salvatore Maugeri, Pavia, Italy; [4]IRCCS Multimedica, Milano, Italy; [5]Università degli Studi di Salerno, Baronissi (SA), Italy

Summary. *This work describes how the application of the Hierarchical Naïve Bayes algorithm to the analysis of GWAS can be deployed on an un-expensive parallel computing infrastructure in the cloud. We evaluate the biological validity of the results obtained by running the algorithm using this novel architecture.*

Introduction

Genome Wide Association Studies (GWAS) are powerful approaches to disentangle the genetic and molecular mechanisms underlying complex traits. The typical approach to these analyses is based on identifying a limited set of associated SNPs to be included into predictive models. These approaches though do not fully capture the genome-wide genetic signature that modulates the probability of developing complex-traits. Recently, we proposed the Hierarchical Naïve Bayes (HNB) [1], an algorithm for predicting the individual level probability of developing a certain disease based on genome-wide SNPs data. Analyzing GWAS data through data mining (DM) algorithms such as HNB is usually a very time-consuming task even when high-performance computational resources are available. To overcome this problem, one of the most common solutions is to exploit parallel computation. This technique allows sharing the computational load among a network of machines, thus dramatically reducing the overall time needed to perform the analysis. As a matter of fact, adapting an existing DM algorithm to be executed in parallel is not a trivial task and the available IT infrastructures to execute parallel computation are very expensive. In this work we propose to deploy HNB on an un-expensive parallel computational infrastructure and we evaluate the biological validity of the results obtained by running the algorithm using this novel architecture.

Methods

For the implementation of the HNB in parallel we exploit the Apache Hadoop framework. This choice enable us to use the MapReduce paradigm to port an algorithm in parallel [2]. Moreover, it offers the possibility to run that algorithm on large clusters of commodity hardware. Hadoop is an opensource Apache Foundation project and is derived from Google's MapReduce and Google File System [2]. In order to build a robust and scalable Hadoop cluster we exploit Amazon Web Services (AWS). Besides fulfilling our needs, AWS also enables us to deploy other related technologies (e.g. temporary storage and network facilities) on demand, with high performance and high IT security standards. Upon the AWS physical IT security layer we deployed a Virtual Private Cloud located in an AWS Region. This enables us to build a logically isolated section of the AWS Cloud where we can launch AWS resources in a virtual network that we define.

Discussion

The parallel implementation of HNB has been tested on a real GWAS dataset represented by 410 long living individuals and a set of 553 genetically matched controls. Results showed that the classification performances obtained using HNB based on a 10 folds Cross Validation on the whole-genome set of variants are significantly higher than those obtained using the same approach but limiting the set of predictors to the top 300 associations (corresponding approximately to $p < 0.001$) selected on each training set (Mean Mattew's Correlation Coefficient (MCC) = 0.24, 95% CI = 0.183-0.297 vs. 0.02, 0.00-0.03). Interestingly, the estimated MCC is significantly higher than 0, corresponding to the condition of no correlation between predicted and real case/control condition. This result suggests that the genetic signature of a complex trait is not represented by few genetic variants passing stringent GWAS significance thresholds, but by the contribution of a genome wide set of variants each showing moderate or null univariate association. We ran simulations on a 5-machines-hadoop cluster in which each node was equipped with 4 virtual CPU and 15GB of RAM. The overall time for the simulation process was around 140 minutes with a cost of 9US$ per simulation, against more than 2 days when HNB was run sequentially using a commercial framework.

References

1. Malovini A et al. BMC Bioinformatics. 2012;13 Suppl 14:S6
2. Dean J, Ghemawat S. Commun. ACM 51, 1, 2008, 107-113.

Integrating Genetic Variants within the i2b2 Framework: the NoSQL way.

Matteo Gabetta[1], Ivan Limongelli[1], Ettore Rizzo[1], Daniele Segagni[2], Riccardo Bellazzi[1]

[1]Biomedical Informatics Labs "Mario Stefanelli", Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy
[2]IRCCS Fondazione Salvatore Maugeri Hospital, Lab of Computer and Systems engineering for Clinical Research, Pavia, Italy

Next generation sequencing data generates large collection of variants that need to be properly organized and queried in databases and datawarehouses. This abstract describes the efforts made in order to integrate, leveraging NoSQL technologies, this information with the phenotypical data managed by the i2b2 framework.

Sample variants are frequently collected in the Variant Calling Format (VCF) as the results of the GATK Unified Genotyper. Often these data pass through a further step of enrichment with tools like those provided by the ANNOVAR pipeline; this tools allow the variants to be integrated with information like: gene names, effects on primary protein structure, frequency in 1000gp and dbSnp, prediction scores, base conservation among species and many others.

Once genetic annotations are available, it is worth to integrate them inside a data warehouse optimized for querying phenotypical data, like i2b2, in order to link these two interconnected aspects.

The reasons for choosing NoSQL technologies, and in particular Apache CouchDB, for managing the genetic data are many, among them we can mention:

- They allow the data model to stay flexible (NoSQL databases are often referred as *schemaless*) so that applications are affected by possible changes in this model considerably less than their SQL-based counterparts.
- The data are stored in a format that is very near to their original one. CouchDB, in particular, stores data into JSON documents (one document per variant in our case); this makes the information stored inside the database easily readable by humans (no need to combine data from different tables). Documents are also easily writable because the only entity to be modified is the document itself.
- Many experiences conducted so far have proven that NoSQL technology is more suitable to scale when the data volume increases [1].

However, NoSQL, and CouchDB in particular, have some limitations compared to traditional SQL databases, the most important being the fact that they do not provide a standard query language. Queries have to be pre-designed and are kept up to date by the database engine; this obviously makes the query process extremely fast but forces the application designer to foresee the axes on which the stored dataset will be queried.

The i2b2 extension we have developed so far is composed by two main parts: the first is the pipeline to populate the DB, which starting from VCF files, runs ANNOVAR, creates the JSONs with ANNOVAR's output and finally stores these files in CouchDB. The JSON structure is based on an object model specifically designed for this task.

The second part is an i2b2 Cell, called NoSQL-NGS Cell, which, along with its plugin for the i2b2 Webclient, allows exploring the variants associated with a Patient Set previously achieved with the i2b2 query process. Despite CouchDB can be queried with REST methods, so that the Webclient could also communicate directly with the database, the presence of a mediator (i.e. the NoSQL-NGS Cell) has proven to be essential because not all the possible client-side queries can be managed with a single query on CouchDB and the assignment to a browser-hosted client of the (potentially) demanding task of aggregating partial results is, in general, a bad practice.

To date the system has been preliminarily tested on standard desktop machine: ranging from a small set of mutations up to the equivalent of 25 exomes (about 550.000 variants), while the set up of the environment scales linearly with the volume of data managed, the query time remains almost instantaneous (<1sec). The whole system has been also deployed on the Amazon Web Services (AWS) environment. Preliminary tests confirm our expectations: once the querying directions have been set, the querying performance and, accordingly, the user experience are very promising.

Next steps include to test the NGS system with a big set of VCF files, to evaluate its performance in real-world tasks, such as managing hundreds exomes; we will also exploit parallelization in several aspects of our system: from the populating pipeline to the CouchDB itself.

[1] Cooper, B. F., Silberstein, A., Tam, E., Ramakrishnan, R., & Sears, R. (2010, June). Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st ACM symposium on Cloud computing* (pp. 143-154). ACM.

Web-based Protein-Interaction Network Analysis Pipeline

Adam Handen[1], M.S. and Madhavi K. Ganapathiraju, Ph.D.[1],
[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Summary
We developed an online network analysis pipeline which is designed to aid biomedical researchers in extracting the protein-protein interaction network of these genes from the human interactome and finding statistically significant associations of these genes with various pathways and diseases. The tool does not require any plugins or downloads, and is freely accessible.

Introduction and Background
Genome-wide association studies identify a collection of genes associated with individual traits; thousands of such studies have now been published; often, the identified genes are poorly characterized, with unknown functions. Interactome based studies would help discover not only functions of individual genes, but also pathway and biological process associations of the trait or disease associated gene sets. Several tools exist to carry out some of the network analyses. PPISURV and ContextNET report interactions across many databases, and other tools like the Correlation Browser can report significant features of datasets; but these tools only report results and tables that can be long and difficult to read. Those tools that do offer visualizations like PINA2 or DTOME can create visuals of networks, but downloading additional plugins or software like Java or Cytoscape. We present here a web based tool that presents easily interpretable visualizations of gene networks and associations with pathways and diseases, without the requirement to install additional software or plugins. It interlinks interactions in the network to fully annotated pages on Wiki-Pi[1], a protein-protein interaction (PPI) web-server.

Methods
The online tool was developed using HTML5 and the D3 Javascript library to produce visualizations. On the server side, PHP and Python perform queries against Wiki-Pi, which contains PPIs and their annotations collected from databases like HPRD, BioGRID, KEGG, Reactome, NHGRI's GWAS catalog, etc. Additional calculations needed for the network analysis are carried out by the python networkx library. The pipeline accepts either one or two lists of genes and performs statistical analyses on them. The genes can be given as gene symbols, Uniprot IDs or Entrez gene IDs. A python script on the server side attempts to find the shortest paths in the human interactome to connect the genes in the first set to the genes in the second set; or if only one gene set is given, it finds the paths to interconnect the genes in that list. The PPIs of individual genes and those that form shortest paths are combined to create the final network visualized. The D3 library provides an interactive graph of the network and labels each gene. The tool links each gene and each PPI in the visualization to its corresponding page on Wiki-Pi. Additionally, users may choose to compute statistically significant of associations of the gene sets: if two gene sets are given, shortest path connectivity between the two sets is presented in comparison to that between random sets of genes. The genes in the interactome are compared against Wiki-Pi's list of known diseases, drugs, pathways, and GWAS. Those with statistically significant overlap are presented with information on the amount of overlap, the p-value for the test, and a representative Venn diagram to better visualize the correlation between gene lists.

Results and Discussion
Our website provides the first stop for interactome based systems biology study of genes identified with GWAS. It presents a visualization of the network and a statistical analysis of the interactome of the genes to determine its association to specific pathways and its overlap with other disease associations. Most importantly, the tool requires no additional software. By providing this tool freely online, we hope to improve the efficiency of researchers who wish to find new genes of interest or find new associations with their existing gene sets. This is ongoing work; we are adding further network analyses and we plan to integrate the tool into Wiki-Pi in the near future. Contact authors for a link to the website.

References

1 Orii, N. & Ganapathiraju, M. K. Wiki-pi: a web-server of annotated human protein-protein interactions to aid in discovery of protein function. *PloS one* **7**, e49029, doi:10.1371/journal.pone.0049029 (2012).

Text mining assisted pathway curation - Full text articles versus abstracts

Ravikumar K.E., PhD, Kavishvar B. Wagholikar and Hongfang Liu, PhD
Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN

Abstract
Annotation of biological pathway databases is largely driven by manual effort with little assistance from text mining. It is a great challenge to the pathway curators to keep up with the pace of ever-growing literature. There have been recent efforts to fill this gap through text mining by identifying the relevant papers and the textual evidence pertaining to pathway information. In the current work, we propose a text mining system that extracts events in molecular pathways from both abstracts and full text articles and its role in assisting manual curation of pathway databases. We specifically investigated the merits of mining full text articles for pathway curation by comparing the performance of our system on both full text articles and biomedical abstracts. From the preliminary results we observed nearly 6% increase in the recall and 5% drop in the precision when comparing the annotations extracted from full text articles against the one from the abstracts. Preliminary analysis on selected pathways from PharmGKB suggest that pathway curators do use their biological knowledge to infer new information that go beyond what is often expressed in wither the full text articles or abstracts.

Description
Despite more than a decade long research in biomedical text mining their role in curation of biological pathway databases and specifically the pathway databases have been very limited. There have been prior efforts [1] to bridge this gap by identifying the textual evidence for pathway curation. In a recent work [2] we evaluated the performance of our rule based text mining system in extracting the pathway events from the biomedical abstracts cited as literature evidence in PharmGKB [3] database. The overall F-measure in recovering pathway events was 34.99%. Majority of the work still focus on only biomedical abstracts while the curators rely on full text articles to annotate pathway databases. In this preliminary work we extended our study by mining pathway events from full text articles related to 4 pathways from PharmGKB.

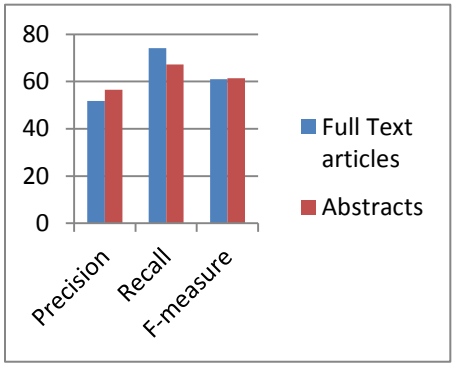


Figure 1 - Evaluation of events extraction on full text articles versus abstracts

We processed 34 full text articles cited as literature evidence for annotations in 4 pathways from PharmGKB. There were 58 total events in PharmGKB which we considered as gold standard annotation. The system extracted 83 unique events from the abstract and the results and discussion section of full text articles. After manual evaluation we found only 43 to be correct leading to precision, recall and F-measure of 51.80%, 74.13% and 60.98% respectively. Figure 1 compares the performance of event extraction of our system between just the abstracts and the full text articles. We observe a slight drop in the overall F-measure despite recall being significant. While mining full text articles did fetch additional events when compared to full text articles we found that there were 30% of events which our system were not able to identify. Error analyses indicate that annotations of pathway information in PharmGKB go beyond what is conveyed in full text articles. The curators who are primarily biologists do lot of inferencing based on their domain knowledge which often reflects in the annotation. This study emphasize the fact that in order to further improve the state of the art of text mining there is an urgent need to integrate background knowledge in biology encoded in ontologies into its workflow.

References

1) Schmidt CJ, Sun L, Arighi CN, et al. Pathway curation: Application of text-mining tools eGIFT and RLIMS-P. Bioinformatics and Biomedicine Workshops (BIBMW), 2012 IEEE International Conference on; 2012: IEEE; 2012. p. 523-8.
2) Ravikumar, K.E., Wagholikar, K B. and Liu H. *Towards pathway curation through Literature mining – A case study with PharmGKB*. (Accepted) *Pacific symposium of Biocomputing* 2014.
3) Hewett M, Oliver DE, Rubin DL, et al. *PharmGKB: the pharmacogenetics knowledge base*. Nucleic acids research. 2002;**30**(1):163-5.

Converting Peoplesoft Data into Semantic Data and RDF

Alexander Loiacono, BS, Nicholas Rejack, MS, Christopher P. Barnes, Michael Conlon, PhD, Erik Schmidt, MS, University of Florida, Gainesville, FL, USA

Summary: The University of Florida runs a VIVO (http://www.vivoweb.org/) implementation to display an assortment of semantic data about research and researchers within our institution. The utility of a VIVO system improves as more data is added to the system. We have developed software to gather data from multiple disparate sources, both on the web and internal to the University. Our software merges this data together in meaningful 'buckets' and joins the data to objects within the VIVO system, which is then able to represent the collection of data semantically. This collection of semantic data is then represented on the web, in the VIVO system, where it can be downloaded and processed by other systems, or simply viewed by an individual within a web browser.

Introduction: An essential feature of the web is that links can break. Because of this, maintaining referential data integrity becomes a priority. When it comes to enterprise systems, however, this feature becomes less desirable especially when implementing semantic systems that rely on URI's as dereferencable identifiers. Institutions seeking to maintain semantic systems with persistent identifiers while adding new data must therefore come up with unique solutions to align new data with existing data. One such semantic system is called VIVO. VIVO is a researcher networking tool at the University of Florida containing over one hundred fifty thousand people and over one hundred thousand other entries including, but not limited to, organizations, research, and courses, all with unique URIs. Gathering all of these people, properly classifying them, and then making the necessary connections (e.g. faculty that has taught the following courses and has the following contact information) is a daunting task.

Methods: The process of maintaining data integrity has been largely automated on several fronts thanks to custom built 'Person Ingest' software and database management tools. Each week, our Person Ingest software runs and combines data coming from multiple data sources to generate people add/subtract RDF files that are then uploaded to the VIVO site. Combining different data sources across multiple platforms is the first step in the process. Operational institutional data is combined with informational institutional data, generating various files for ingest. The first is called 'contact_data' and contains all pertinent information about an individual such as email, phone, title, and department. Next, a 'position_data' file is generated which contains information such as start and end dates, position, and job codes. Finally, a 'privacy_data' file is generated that marks whether a person should have their information published in VIVO or hidden from public view. All data for ingests is pulled from multiple authoritative tables across multiple platforms on a nightly basis. Live SPARQL queries pull realtime data from VIVO that the incoming data is then aligned with. Data such as department ID's and names are used for linking on the production site. As part of this process, certain entities belonging to protected departments or organizations (such as the University Police Department) have their contact data removed before publishing to VIVO websites. All of these data sources, in .CSV format, are fed to the ingest software that generates an add and subtract RDF file. This process takes about an hour for all the data to be read in, organized, processed, and output into the RDF files. After the RDF files are generated, they are uploaded to VIVO via the "Add or Remove RDF Data" page. The add file upload time is linear to the size of the file and usually takes anywhere from 5 30 minutes. The same can be said of the sub file, though it is generally smaller and can sometimes be completed in a shorter time frame. At this time, the RDF upload process is not fully automated as an administrative user needs to manually start the upload within the VIVO GUI. We feel this lack of full automation is beneficial, as it provides a window for manual data inspection before it is published to the production website.

Results: After many months of testing for sound methodology and acceptable data integrity, the Person Ingest software has been vetted for inclusion in our weekly maintenance process on the University's production VIVO systems. To maintain vigilance, we have since implemented several additional pieces of software that check VIVO data integrity and report back on any discrepancies found. Examples of this would be person objects with a home department of 'University Police Department' or authorships not properly linked to authors.

Discussion: The ingest software is extensible and can be used to harvest data for other areas of a VIVO site. The next step in our process is to automate the addition and linking of courses being taught at the University of Florida. This poster will discuss ways to extend this method of operation to other institutional data, regardless of the institution or data source.

CODE: A Shared Computing Environment for Precision Medicine

Subha Madhavan PhD[1], Varun Singh MS[1], Lei Song MS[1], Anas Belouali MS[1], Krithika Bhuvaneshwar MS[1], Yuriy GusevPhD[1], Peter McGarvey PhD[1], Michael Harris MA[1]

[1]Innovation Center for Biomedical Informatics, Georgetown University, Washington DC

Abstract: The new and emerging field of Systems Medicine, an application of systems biology approaches to biomedical problems in the clinical setting, leverages complex computational tools and high dimensional large datasets to derive personalized assessments of disease progression, risk, and drug response. The Clinical - Omics Discovery Environment (CODE) is a flexible web-based data framework that serves to enable basic, translational, and clinical research by integrating patient characteristics and clinical outcome data with a wide variety of large-scale high throughput research data in a unified environment to drive hypothesis generation and validation of molecular markers. CODE provides a variety of disease-centric analyses and promotes the use of informatics tools by physician scientists and basic researchers for large omics datasets in the context of clear clinical endpoints.

Introduction: We live in a data intensive era where the cost of genome sequencing is expected to drop to $100/genome in the next decade and the capacity to sequence a billion people will have been realized in the next twenty years requiring 3000 PB of storage. A number of public and private projects are already contributing to this biological data deluge. We present a shared computing collaboration framework: Clinical - Omics Discovery Environment (CODE) that provides a novel solution to address challenges in applying big data omics to precision medicine. CODE powers disease-specific or function-specific data portals using a robust and well-documented application program interface (API) to help address challenges of big data, integration, analysis, and interpretation.

Methods: The architectural framework for CODE has been functional for the past 2.5 years. We support over 350 users, biospecimen from over 8000 patients along with linked clinical and omics data, and a variety of cell lines. CODE provides a flexible object model for clinical attributes, allowing the integration of new data types very quickly. Dynamic explorations of data are made possible through the CODE web interface that provides users with a comprehensive set of analysis routines and visualizations for a rich user experience. The application is written in Groovy & Grails, an open source application framework that runs on the Java Virtual Machine. The jQuery JavaScript library is used to provide a cohesive and interactive interface. For more complex data visualization, the Adobe Flex framework provides capabilities to handle complex charts. Besides the components developed in-house, CODE also incorporates many third-party tools that provide data visualization capabilities. For example, Java TreeView is used to display heat maps, Cytoscape to display interaction networks, and JBrowse provides a genome browser with multiple annotation tracks. CODE can be deployed locally or on cloud computing services.

Results & Discussion: CODE-powered data portals are supporting a variety of user groups in multiple domains to enable hypothesis generation and translational medicine. CODE tools have been leveraged to support numerous studies including: the detection of prognostic markers for relapse in colorectal cancer samples; a vaccine safety study to curate and identify proteins that link vaccine ingredients to autoimmune diseases (**Figure 1**); and a study to detect key metabolites related to disease severity, progression, and responsiveness to corticosteroid treatment in children with Duchene Muscular Dystrophy.

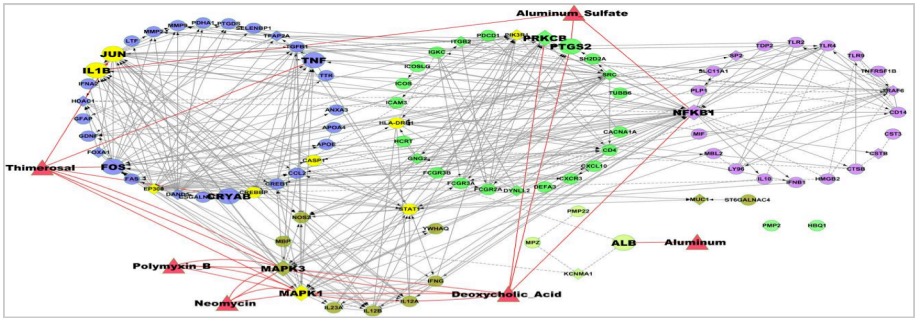


Figure 1. Network of Guillain-Barre syndrome & vaccine ingredients

Conclusion: The long-term vision for CODE involves establishing a robust and comprehensive systems medicine platform that can directly impact health care in collaboration with hospital networks by providing more effective clinical decision support using multi-omics data. Future development will focus on integration with electronic health records (from collaborating hospitals) based on its expected impact on both clinical research and clinical practice.

Integrative analysis of genomic data to identify candidate biomarkers of resistance to platinum-based chemotherapy in ovarian cancer

Sheida Nabavi[1], Mayinuer Maitituoheti[1,2], Ewa Przybytkowski[3], Dennis Wall[1,2], Mark Basik[3] , and Peter Tonellato[1,2]

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, MA; [2]Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA; [3]Lady Davis Institute for Medical Research, McGill University, Montreal, QC, Canada

Abstract

Drug resistance is one of the major challenges in the treatment of ovarian cancer. To facilitate identification of candidate biomarkers of response to platinum-based chemotherapy in ovarian cancer we analyzed gene expression, somatic mutation and copy number aberration data of platinum sensitive and resistant samples from TCGA.

Introduction

Ovarian cancer is the deadliest and the second most common gynecologic cancer. Almost all women diagnosed with ovarian cancer receive combination of cytoreductive surgery and platinum-based chemotherapy. Although many of patients respond to the platinum-based chemotherapy, the majority does not respond and will ultimately succumb to the disease. Therefore, drug resistance is an urgent problem in the current treatment for ovarian cancer; and finding a clinical practical way to identify in which individuals' platinum-based chemotherapy will be effective is essential to obtain a consistent outcome and meaningful benefit, as well as to avoid unnecessary toxicity and cost of healthcare.

The objective of this work is to facilitate identification of candidate biomarkers of response to platinum-based chemotherapy in ovarian cancer using computational approaches. To accomplish this, we conducted comparative genomic analysis of platinum sensitive and resistant samples from the cancer genome atlas (TCGA) and applied integrative analysis of genomic data. The goal is to find candidate genes for resistance to platinum as potential biomarkers. It is hypothesized that candidate genes involved in resistance are among copy number aberrant and mutated genes and that the expression patterns of those genes match the mutation and copy number patterns. Therefore in this work we looked for genes that are differentially mutated (amplification/deletion and/or point mutation) while are differentially expressed in the platinum resistant versus platinum sensitive samples.

Methods

We used processed genomic data for ovarian cancer, which is publically available from TCGA together with pertinent clinical information. Using the clinical information, we selected the data for 97 platinum resistant and 234 platinum sensitive primary tumors. For the genomic comparative analysis we integrated gene expression, copy number alteration and somatic mutation datasets of these samples.

We pursued two approaches. In the first approach we applied genomic comparative analysis first and then applied integrative analysis. In this approach, we compared the mutation status (amplification/deletion and/or point mutation) and expression values of all genes for the resistant versus sensitive samples separately; and then integrated the differential mutation and differential expression information to obtain a short list of the candidate genes. For the integrative analysis, we also calculated the Pearson correlation between expression values and copy number values of differentially mutated and differentially expressed genes. In the second approach, we applied integrative analysis first and then we applied comparative analysis. In this approach, we integrated mutation and gene expression data for resistant and sensitive samples separately to find the candidate genes for each group; and then compared the two gene lists to obtain a short list of genes that are significant in one group but not in the other group. In this approach, we used module network and Bayesian analysis as in CONEXIC(1) for the integrative analysis. Then, we compared the results of the above two approaches and selected candidate genes that showed up in the both approaches as a final result. The final result is a short list of genes, which are mutated and expressed differentially in resistant versus sensitive tumors. This list is explored further for its association with signaling pathways.

References

1.Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell 2010; 143(6); 1005-17.

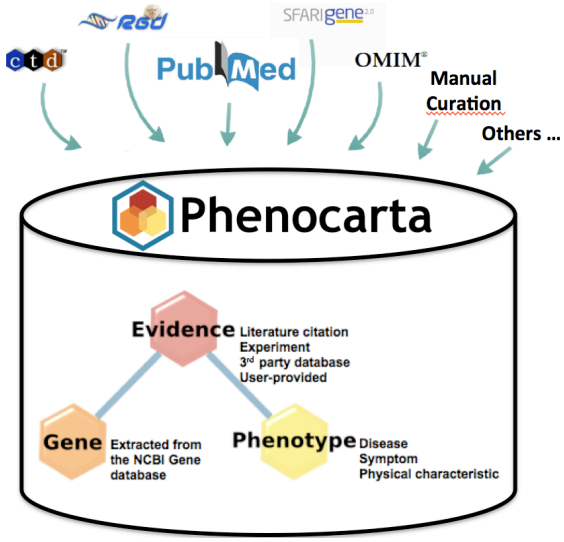
Phenocarta: A Comprehensive Gene-Disease Database for the Interpretation of Genomics Studies

Elodie Portales-Casamar, PhD[1], Nicolas St-Georges, MSc[1], Paul Pavlidis, PhD[1]
[1]Centre for High-Throughput Biology, University of British Columbia, Vancouver, BC, Canada

Summary.
Phenocarta is a widely applicable database designed for research applications in genetics, genomics and proteomics. The system is unique in coverage and includes novel approaches to annotating and sharing gene-disease relationships. Its development involves addressing interesting issues in data quality, provenance and redundancy. We are eager to engage the research community to further develop and apply the system.

Introduction and Background.

Understanding the genetic basis of diseases is key to the development of better diagnoses and treatments. Unfortunately, only a small fraction of the existing data linking genes to diseases is available through online public resources and, when available, it is scattered across multiple access tools. Phenocarta (http://phenocarta.chibi.ubc.ca) is a knowledgebase that consolidates information on genes and diseases across multiple resources and allows tracking and exploring of the associations. Diseases are recorded using controlled vocabularies such as the Disease Ontology to facilitate computational inference and linking to external data sources. We seek to enhance the usability of Phenocarta through the addition of new features and more data.



Methods, Results, and Discussion.

When Phenocarta was published in February 2013 (a.k.a. Neurocarta), it contained 30,000 lines of evidence linking over 7,000 genes to 2,000 different diseases. It has now been expanded to about 142,000 lines of evidence linking about 12,000 genes to almost 3,000 diseases. These additional annotations come from updates from the external resources we had already integrated, as well as the inclusion of new resources such as the NIH GWAS catalog (http://www.genome.gov/26525384) or the Disease Ontology Annotation Framework (http://doa.nubic.northwestern.edu/pages/search.php). Another data source that we are currently incorporating into Phenocarta is meta-analysis of gene expression. We have developed a tool to perform such analysis through the Gemma framework (http://gemma.chibi.ubc.ca) and are now investigating the best way to incorporate this evidence type into Phenocarta.

One of the challenges of using diverse data such as animal models, gene expression analysis and human genetics studies of varying power is how the data can be integrated and evaluated fairly. To standardize evidence quality information in Phenocarta, we have developed a five-tier system associated with carefully documented criteria. However, manually annotating all lines of evidence with these codes is not feasible, so we are developing and evaluating automated approaches. To this end we have created a manually-curated gold standard set to use for algorithm training and evaluation. The most naive method for automated assignment is to assign scores simply based on the overall quality of the resource the annotation comes from, ignoring the potential variability within the resource. We are currently testing approaches based on features of the structured and free-text information in the publications (e.g., MeSH headings) associated with a line of evidence and will present preliminary results based on cross-validation with the gold-standard showing improved performance over the naive scheme. The coarse granularity of the proposed scoring scheme is purposeful, because at some level quality is subjective. The real power of this approach will be seen when evidence is integrated across sources.

In conclusion, we believe that the new data and features we are incorporating into Phenocarta will drastically change the way investigators can use gene-disease information in the interpretation of genomics data.

Developing Sharable CDS Rules Based On Pharmacogenomics Guidelines Using The Health eDecisions Interoperability Standard

Davide Sottara, PhD[1], Robert R. Freimuth, PhD[2]
[1]Department of Biomedical Informatics, Arizona State University, Scottsdale, AZ
[2]Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract:

Pharmacogenomics (PGx) clinical guidelines are published to provide recommendations for genetic test results. Currently, these guidelines are distributed in an unstructured format designed for human consumption. We propose the adoption of a formal representation to reduce ambiguity and facilitate transformation into clinical decision support (CDS) rules. We adopted the Health eDecisions (HeD) format to represent PGx guidelines. We report our evaluation and integration of HeD with PGx-specific data models and terminologies.

Description:

Pharmacogenomics (PGx) is the study of how an individual's genetic makeup affects their response to drugs. As clinical trials and translational studies advance PGx knowledge, recommendations are being developed by groups such as the Clinical Pharmacogenetics Implementation Consortium (CPIC)[1] to help caregivers interpret and apply genetic data at the point of care.

Given the high degree of specialized knowledge, our rapidly evolving understanding of molecular systems, and the potential for impact on patient care, medical insitutions are developing computer-based clinical decision support (CDS) systems to help physicians apply PGx guidelines in practice. Currently, however, no commercially available electronic health record natively supports PGx-based CDS. To lower the barrier to adoption and facilitate the integration of PGx into clinical practice, we are evaluating and applying state-of-the-art techniques in guideline modelling and formalization to create machine-readable representations that can be consumed and shared between both humans and clinical information systems.

The Health eDecisions (HeD) XML interchange format[2] models actionable CDS interventions, such as rules and order sets, that can be extracted directly from guidelines. HeD is sponsored by the Office of the National Coordinator for Health Information Technology (ONC) and balloted as an HL7 standard. It is designed to facilitate the large scale distribution and integration of CDS. While HeD has not been explicitly designed for PGx, the schema is flexbile and extensible in terms of the vocabularies, data models and logic expressions it can support.

We adopted HeD for its structured format and rich expression language. The HL7 virtual medical record model was used to represent common clinical concepts, such as laboratory test results, and the RxNorm vocabulary was used to represent drugs. The Pharmacogenomic Guideline Model, a domain-specific model that was developed based on PGx guidelines[3], was used to represent concepts such as genes, variant alleles, and molecular phenotypes.

For our initial analysis, two of the 10 CPIC guidelines that have been published to date were selected for full representation using the HeD schema. Specifically, the HLA-B/abacavir and TPMT/thiopurines guidelines were selected because the genetic test results are represented differently. The HeD schema, as well as the supporting models and terminologies, were evaluated with respect to their ability to represent the semantics of the PGx guidelines. Limitations were identified in all components and were addressed where possible.

These results demonstrate the feasibility of expressing PGx guidelines using the emergent HeD standard. Computable representations of PGx knowledge and the subsequent sharing of CDS rules will lower barriers to the clinical adoption of PGx by increasing scalability and decreasing the cost of implementation and maintenance. Ongoing and future work includes the adoption of the structured format to catalogue the guidelines and deploy them in a runtime environment to create a PGx decision support service.

References

1. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. Clin Pharmacol Ther. 2011 Mar; 89(3):464-7.
2. Health eDecisions [Internet]. [cited 2013 October 10]. Available from: http://wiki.siframework.org/Health+eDecisions+Homepage
3. Freimuth RR, Chute CG. Pharmacogenomic Drug Dosing Guidelines: A Model to Support Adoption. 2013 Joint Summits on Translational Science. San Francisco, CA. March 19, 2013.

Predicting synergistic therapies for triple negative breast cancer through an integrated network model

Francesca Vitali MD[1], Francesca Mulas PhD[1], Alberto Zambelli [2], Riccardo Bellazzi PhD[1]
[1]Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Italia;
[2]Fondazione Salvatore Maugeri, Pavia, Italia

Abstract

Current research indicates that biological networks might offer conceptual frameworks that could elucidate pharmacological strategies. We developed a computational method that predicts target combinations by formalizing the ranking of network nodes. The application to breast cancer highlights potential targets with a synergistic effect that may be proposed as novel treatments.

Introduction

In complex disease, such as cancer, single drug therapies are often found to generate side-effects and interactions with other drugs, thus hampering the final therapeutic success. For this reason, current research focuses on assessing the drug treatments as a whole rather than considering them individually. Based on this idea, polypharmacology aims to achieve superior therapeutic efficacy and safety by designing chemical entities that can simultaneously target different points of a given disease[1]. Among the different types of tumor, Triple Negative Breast Cancer (TNBC) is a heterogeneous and aggressive sub-class whose biology is poorly understood[2]. This type of tumor does not respond to the standard therapies, thus, it is reasonable to plan a synergistic strategy for the TNBC based on its unique pathway characteristics. In this context, a network-centric modeling seems the elective strategy to deal with multicomponent therapeutics in complex diseases, as it "naturally" offers new therapeutic views and recommendations for drug repositioning[1]. We developed a network-based approach that, thanks to the integration of different data sources, could better define the global picture of the TNBC disease status with the aim to select multi-target therapies.

Methods

In this work we propose an adapted application to the case of TNBC cancers of a novel methodology developed in our previous work[1]. This approach focuses on disease-specific protein networks to rank new target candidates thanks to the definition of the *Topological Score of Drug Synergy* (TSDS). The protein interaction network for the TNBC subtype is constructed starting from the genetic changes related to the disease and by integrating them with the available knowledge on the this disease. Afterwards the method provides the selection of two subsets of network proteins: the *disease nodes* and the *target nodes,* corresponding to the proteins that need a therapeutic effect and to the nodes sources of this effect, respectively. The topological analysis of the network allowed us to define a score system, the TSDS, aimed at reducing the search space to determine the most promising combinations for experimental evaluation. The TSDS performs the ranking of the best target combinations considering the target positions with respect to the disease nodes. Thanks to the application of this system score we might identified the significant multi-target candidates and provided a general framework for a synergistic therapy.

Results and discussion

The analysis of the best target combinations demonstrates that the method could elucidate the interactions of the complex disease under study and automatically suggest potential drug interventions. This work implemented a computational platform that integrates pathways, protein-protein interactions, transcriptional analysis to result in a comprehensive network for new multi-target discovery that can be applied to every complex diseases. Higher integration across different system biology platforms could provide mutual-support and validate data as well as speed up multi-target drug discovery.

References

1. Vitali F, Mulas F, Marini P, Bellazzi R. Network-based target ranking for polypharmacological therapies. JBiomedInform. 2013; 46: 876-881.
2. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao, Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012 486(7403): 395-399.

Automated Physician Order Recommendations and Outcome Predictions by Data-Mining Electronic Medical Records

Jonathan H. Chen, MD, PhD[1], Russ B. Altman, MD, PhD[1,2*]
1 Department of Medicine, Stanford University, Stanford, CA 94305, USA.
2 Departments of Bioengineering and Genetics, Stanford University, Stanford, CA 94305, USA.
*To whom correspondence should be addressed. E-mail: russ.altman@stanford.edu

Abstract

The meaningful use of electronic medical records (EMR) will come from effective clinical decision support (CDS) applied to physician orders, the concrete manifestation of clinical decision making. CDS development is currently limited by a top-down approach, requiring manual production and limited end-user awareness. A statistical data-mining alternative automatically extracts expertise as association statistics from structured EMR data (>5.4M data elements from >19K inpatient encounters). This powers an order recommendation system analogous to commercial systems (e.g., Amazon.com's "Customers who bought this…"). Compared to a standard benchmark, the association method improves order prediction precision from 26% to 37% ($p<0.01$). Introducing an inverse frequency weighted recall metric demonstrates a quantifiable improvement from 3% to 17% ($p<0.01$) in recommending more specifically relevant orders. The system also predicts clinical outcomes, such as 30 day mortality and 1 week ICU intervention, with ROC AUC of 0.88 and 0.78 respectively, comparable to state-of-the-art prognosis scores.

Introduction

Electronic medical records (EMR) can improve patient safety and healthcare cost efficiency, but that depends on meaningful use of the data[1]. This will require effective clinical decision support (CDS) content, particularly to drive clinical orders (labs, imaging, medications, etc.), the concrete manifestation of clinical decision making. Order sets, risk scores, and similar CDS constructs help reinforce consistency and compliance with best-practices[2,3], but their conventional development is limited by a top-down approach. This approach requires manual production of CDS content, feasible for only a limited number of common scenarios, and often with limited end-user awareness[4]. With the progressive digitization of clinical data in EMRs, a Big Data[5,6] approach can instead crowd-source clinical expertise from the bottom-up by data-mining EMRs. Such an approach could continuously "learn" in real-time by streaming in accumulating EMRecords into data-driven models of clinical expertise, even as it is simultaneously applied to patient care with direct EMR integration.

Background

Prior work in automated CDS content development includes association rules and Bayesian networks between orders and diagnoses, and review of possible order set and corollary order content by subject experts[7–10]. With inspiration from analogous problems of information retrieval in recommender systems, collaborative filtering, market basket analysis, and natural language processing, we initiated an item association order recommendation framework[11] analogous to Netflix or Amazon.com's "Customer's who bought A also bought B" system[12]. Here we update our initial efforts with a much larger dataset that includes non-order data to better define a patient's clinical context, propose an alternative evaluation metric to identify recommendation methods that highlight items specifically relevant to a given clinical scenario, and use the framework to predict clinical outcomes.

Methods

Deidentified, structured patient data from inpatient hospitalizations at Stanford University Hospital in 2011 was extracted by the STRIDE project[13]. Extracted data covers patient encounters starting from their initial (emergency room) presentation until hospital discharge. With >19K distinct patients, the data consists of >5.4M instances of >17K distinct clinical items, with patients, instances, and items respectively analogous to documents, words, and vocabulary items. The clinical items include >3,500 medication, >1,000 laboratory, >800 imaging, and >700 nursing orders. Non-order items include >1,000 lab results, >5,800 problem list entries, >3,400 admission diagnosis ICD9 codes, and patient demographics on age, gender, and date of death. Numerical data was binned into categorical data, particularly lab results, based on "abnormal" flags as established by the clinical laboratory. The ICD9 coding hierarchy was collapsed as necessary into diagnosis codes with a significant number of instances.

The relationship between item instances covered and the top clinical items considered is consistent with the "80/20 rule" in the form of a power law distribution[14]. This property allows one to ignore most clinical items with minimal information loss. In this case, ignoring sparsely populated clinical items with <256 instances (0.005% of

all instances) reduces the effective item count from >17K to 1.5K (9%), while only reducing item instance coverage from 5.4M to 5.1M (94%). Computational efficiency of subsequent order recommendations improves significantly with this simplification, given methods requiring $O(m^2)$ space and $O(q * m \log m)$ time complexity, where m is the number of clinical items considered and q is the number of query items for a recommendation.

A pre-computation step collects frequency statistics on clinical item instance co-occurrences from a training set of 16,408 randomly selected patients to build an item association matrix, based on the definitions in Table 1. These statistics drive subsequent recommendations by approximating Bayesian conditional probabilities as in Table 2.

Notation	Definition
$\mathbf{n_A}$	Number of occurrences of order A
$\mathbf{n_{ABt}}$	Number of occurrences of order B following an order A within time t
N	Total number of patients

Table 1 - Pre-computed frequency statistics for clinical items. Counting repeats allowed.

Probability	Estimate	Notation / Notes
P(A)	n_A / N	BaselineFreq(A)
P(AB)	n_{AB} / N	n_{AB} ("Support") only counts directed association where A occurs *before* B
P(B|A) = P(AB) / P(A)	n_{AB} / n_A	ConditionalFreq(B|A) ("Confidence") Frequency of B, given A
P(B|A) / P(B) = P(AB) / P(A)*P(B)	(n_{AB}/n_A) / (n_B/N)	FreqRatio(B|A). Estimates likelihood ratio. Expect = 1, if A and B occur independently

Table 2 - Bayesian probability estimates based on item frequency statistics.

To generate order recommendations from the above association statistics, query clinical items ($A_1,\ldots,A_q$) are used to select item association pairs from the pre-computed association matrix for all possible target orders ($B_1,\ldots,B_m$). Target orders are ranked by a score such as ConditionalFreq($B_j|A_i$), the maximum likelihood estimator for the probability of order B_j occurring after query item A_i. As previously noted[11], ranking by ConditionalFreq identifies likely orders, but also tends to yield non-specific orders (e.g., CBC, IV saline) that are common overall, yet not necessarily "interesting." To identify orders more significantly relevant to the query, recommendations are ranked or filtered by FreqRatio(B|A), comparable to the TF*IDF (term frequency * inverse document frequency) information retrieval concept[15].

To quantify the significance of item associations, *-2 log FreqRatio* can approximate a chi-square statistic[15] or the chi-square statistic can be directly calculated by comparing observed vs. expected occurrence counts. Issues with misinterpreting association strengths in the setting of inadequate data (heuristics advise at least 5 occurrences to be reliable[15]), are mitigated by excluding rare items occurring <0.005% of the time as previously described.

Given q query items, the above method generates q scored lists of all m possible orders. These are aggregated into a single scored recommendation list by taking a weighted average of the component scores, weighted inversely proportional to their respective query item baseline frequencies (lending more weight to less common, more specific query items). Unweighted score averaging and a Naïve Bayes[15] style composite product of the component conditional probabilities (i.e., conditional frequencies) were also attempted, though the weighted average method was retained as it yielded the best results.

While there is no well accepted notion of recommendation quality, accuracy in predicting subsequent items is the most commonly measured, with precision (positive predictive value) and recall (sensitivity) correlating with end-user satisfaction[16]. A test set of 1,903 patients was randomly selected, separate from the training set. For each test patient, all clinical items from the first 4 hours of their hospital encounter were used (average of 29) to query for 10 recommended orders that were compared against the actual subsequent orders within the first 24 hours (average of 15). To quantitatively recognize recommenders that yield results that are more meaningfully relevant to a query and not simply common, we introduce the alternative metrics of inverse frequency weighted precision and recall, based on the following function definition: *TP(i) = {1 if recommended item i is a true positive, 0 if not}*. Likewise *FP(i)* for false positives and *FN(i)* for false negatives. The inverse frequency weighted precision and recall metrics are defined below in summation notation, with components weighted by the inverse baseline frequency of each item *i* (n_i/N). Note that the common constant factor N can be cancelled out to yield:

$$\text{Weighted Precision} = \Sigma\,(1/n_i)*TP(i) \,/\, (\,\Sigma\,(1/n_i)*TP(i) + \Sigma\,(1/n_i)*FP(i)\,)$$

$$\text{Weighted Recall} = \Sigma\,(1/n_i)*TP(i) \,/\, (\,\Sigma\,(1/n_i)*TP(i) + \Sigma\,(1/n_i)*FN(i)\,)$$

The association framework was also applied towards "recommending" non-order items to predict outcomes such as patient death and ICU intervention. For the latter, a composite "AnyICU" clinical item was defined as the

occurrence of interventions including mechanical ventilation, vasopressor infusion (epinephrine, norepinephrine, dopamine, phenylephrine, vasopressin, dobutamine), or continuous renal replacement therapy (CRRT). Taking 1,905 test patients separate from the training set, their first 24 hours of clinical items were used to query the association model for the probability (ConditionalFreq(B|A)$_t$) of an outcome event within *t* time (30 days for death, 1 week for AnyICU) and compared them vs. actual event rates by receiver operating characteristic (ROC) analysis.

Results

Table 3 illustrates example order recommendations. Table 4 reports accuracy metrics for different recommendation methods, illustrating the trends toward the best results. Table 5 reports the ROC area-under-curve (AUC) prediction accuracy for outcomes of 30 day mortality and 1 week use of AnyICU. Table 6 illustrates an inverted query example, identifying items commonly *preceding* an outcome event.

		Frequency / Likelihood			
Rank	**Description**	**Conditional**	**Baseline**	**Ratio**	**p**
1	TYPE AND SCREEN	0.98	0.78	1.3	0.00
2	Pantoprazole (Intravenous)	0.75	0.42	1.8	0.00
3	TRANSFUSE RBC	0.55	0.52	1.1	0.20
4	PANTOPRAZOLE IV INFUSION	0.51	0.03	16.0	0.00
5	CONSULT MEDICINE	0.32	0.16	2.0	0.00
6	LIPASE	0.29	0.26	1.1	0.15
7	ISTAT TROPONIN I	0.28	0.28	1.0	0.96
8	CONSULT GASTROENTEROLOGY	0.22	0.03	8.6	0.00
9	UPPER GI ENDOSCOPY	0.21	0.08	2.8	0.00
10	ISTAT, VBG AND LACTATE	0.21	0.19	1.1	0.47
11	Oral Electrolyte Solution (Bowel Prep)	0.17	0.03	5.3	0.00
12	OCTREOTIDE INFUSION	0.17	0.01	11.7	0.00
13	TRANSFUSE FFP	0.16	0.16	1.0	0.91
14	Benzocaine+Tetracaine (Topical)	0.09	0.04	2.0	0.00
15	H. PYLORI AG, STOOL	0.08	0.02	4.9	0.00

Table 3 – Example orders recommended when query by admitting diagnosis of GI Hemorrhage, ranked by ConditionalFreq(B|A)$_{day}$ and filtering out those with FreqRatio(B|A)$_{day}$ <1. Example interpretation: Given a GI Hemorrhage, 75% of patients receive IV Pantoprazole (standard initial treatment for an acute GI bleed) within 24 hours. This is somewhat more likely (FreqRatio 1.8) than for all patients in general, though even the baseline of 42% is relatively common as IV Pantoprazole is used for non-GI bleed scenarios (e.g., prophylaxis against stress ulcers). For comparison, the Pantoprazole IV continuous infusion is less common (51%), but has a higher relative likelihood (freqRatio 16.0), as it is used almost exclusively in the treatment of GI bleeds.

Ranking Method	**Time Span**	**Ratio Filter**	**Recall**	**Precision**	**F1-Score**	**Weighted Recall**	**Weighted Precision**	**Weighted F1-Score**
Random			1%	2%	1%	1%	1%	1%
BaselineFreq*			17%	26%	19%	3%	24%	4%
ConditionalFreq	Any		22%	31%	23%	5%	29%	6%
ConditionalFreq	Hour		19%	27%*	20%	5%	17%	6%
ConditionalFreq	Day		**27%**	**37%**	**28%**	7%	**37%**	9%
ConditionalFreq	Day	Yes	9%	17%	11%	15%	14%	**12%**
FreqRatio	Day		8%	12%	9%	**17%**	8%	10%

Table 4 – Average accuracy statistics for recommendation methods across 1,903 test patients comparing 10 system recommended orders vs. actual orders occurring within 24 hours. The ConditionalFreq ranked methods are subdivided by what time span *t* that their item association counting accepts. The last pair of methods use the FreqRatio for filtering (excluding recommendations with FreqRatio <1) or ranking. Bolded entries represent the best value for each metric. *All metrics are compared against the BaselineFreq method as a benchmark, with all yielding p<0.01, except precision of the ConditionalFreq (1 Hour) method, having p = 0.08.

	Death	**Any ICU**
Evaluation period	30 days	1 week
Patients screened	1,905	1,905
Patients evaluated, excluding those with outcome occurring during 24 hour query period	1,898	1,765
Patients with outcome subsequently occurring during evaluation period	44 (2.3%)	55 (3.1%)
ROC AUC score for association prediction	0.88	0.78

Table 5 – ROC area-under-curve prediction metrics for 30 day mortality and 1 week requirement for ICU intervention (ventilator, vasopressor infusion, CRRT) based upon 1,905 test patients' first 24 hours of query clinical items.

Rank	Description	Frequency / Likelihood Conditional	Baseline	Ratio	p
1	COMFORT CARE MEASURES	0.11	0.02	5.22	0.00
2	LIBERALIZE VISITATION POLICY	0.08	0.02	5.09	0.00
3	LACTIC ACID (High)	0.46	0.11	4.12	0.00
4	NOREPINEPHRINE IV INFUSION	0.15	0.04	3.84	0.00
5	CALCIUM CHLORIDE IV INFUSION	0.06	0.01	3.77	0.00
6	Citrate + Sodium Bicarbonate (CRRT)	0.05	0.01	3.68	0.00
7	CONSULT TO PALLIATIVE CARE	0.15	0.04	3.60	0.00
8	OSMOLALITY, SERUM (High)	0.07	0.02	3.55	0.00
9	pH Venous (Low)	0.23	0.06	3.51	0.00
10	LUNG PROTECTIVE VENTILATION	0.07	0.02	3.49	0.00

Table 6 – Inverted query example showing the top "recommendations" for items that occur *prior* to a query item of patient death, ranked by $\text{FreqRatio}(B|A)_{week}$. This recognizes that many deaths are anticipated with a greater likelihood for ordering "Comfort Care Measures" and "Liberalize Visitation Policy," representing reprioritization of care for patients with expected imminent death. Complementary to that are deaths preceded by aggressive life-supporting ICU interventions including vasopressors (norepinephrine), continuous renal replacement therapy (CRRT), and mechanical ventilation for ARDS (lung protective ventilation protocol). Inverse queries can appropriately "recommend" non-order items such as abnormal lab values as well, in this case recognizing that lactic acidosis (high lactic acid) and acidemia (low pH) disproportionately precede death.

Discussion

The item association system developed above, analogous to commercial recommender systems, recommends physician orders and predicts clinical outcomes based on statistics data-mined from electronic medical records. As illustrated in Table 4, personalizing order recommendations with the ConditionalFreq ranking method improves accuracy compared to the standard BaselineFreq benchmark method that only functions as a general "best seller" list, recommending the overall most common orders, irrespective of query items.

Demonstrated again is the importance of temporal information in order recommendation[11], with accuracy optimized when the association time span *t* is comparable to the evaluation time frame. Specifically, when predicting orders occurring within 24 hours of hospitalization, shorter time span filters (e.g., one hour) result in the recommender missing relevant associations for orders outside the filter time, while longer time span filters (e.g., any time) result in the recommender being distracted by associations that occur outside the relevant 24 hour evaluation period. Similarly, when predicting 30 day mortality and 1 week ICU intervention, the time span filters should optimally be adjusted to one month and one week, respectively.

Qualitative examples in Table 3 indicate that FreqRatio based methods can provide more specifically relevant recommendations, but these approaches inherently perform worse by standard accuracy metrics, as confirmed in Table 4. While standard accuracy metrics favor common items, it is more impressive to correctly predict a rare item (e.g., pantoprazole infusion) than the relatively mundane correct prediction of a common item (e.g., Type & Screen). Alternative metrics, the inverted frequency weighted precision and recall, are introduced here to preferentially score prediction of uncommon items. Interestingly, the ConditionalFreq method that performs best on standard accuracy metrics still performs best by the weighted precision metric. It is only for weighted recall that the FreqRatio based methods show improvement (3% to 17%, $p<0.01$). This reinforces the notion that the two approaches serve different purposes and can both be useful depending on the goals of the query.

Table 5 reports the association framework's ability to predict clinical outcomes with ROC AUC of 0.88 for 30 day mortality and 0.78 for requiring ICU intervention within 1 week of hospitalization. These are comparable to state-of-the art prognosis scoring systems such as APACHE, MPM, and SAPS with scores ranging from 0.75 to 0.90 for predicting hospital mortality[17] and CURB-65, PSI, SCAP, and REA-ICU with scores ranging from 0.69 to 0.81 for predicting early ICU admission[18]. Other prediction possibilities could include hospital length of stay, readmissions, and many others, though the virtue of the framework is that it can predict any item labeled as an outcome event with minimal incremental effort in future work.

While the FreqRatio based methods elaborated here help distinguish specifically relevant orders from those that are simply common, a primary concern with this method is favoring common practices that are not actually ideal. With preliminary results on predicting clinical outcomes above, a tempting possibility will link recommendations to favorable outcomes instead of just prevalence, but ultimately this concern will only be proven or disproven by deploying these methods in a prospective clinical trial. Another general concern is that order recommenders may favor over-utilization by encouraging unnecessary orders. The framework can counter-balance this by recommending *against* uncommon orders, and future work will explore personalized prediction of lab result pre-test probabilities to recommend against lab tests unlikely to impact clinical care. Another limitation of the current item association method is that it only considers pair-wise associations, thus querying with multiple items assumes independence between the query items. Incorporating more complex models such as Bayesian networks[8] is

possible, but unclear whether significant accuracy would be gained in exchange for the lost computational efficiency of a simpler model.

In closing, this represents another step in ongoing work towards mature clinical decision support systems that will unlock the Big Data potential of electronic medical records. A clinical order recommendation framework is enhanced here with additional non-order data to better define clinical contexts, reporting of significance statistics for individual recommendations to further aid interpretability, multiple evaluation metrics to discern common from specifically relevant items, and application towards predicting clinical outcomes.

Acknowledgements

Project supported by the Stanford Translational Research and Applied Medicine (TRAM) program in the Department of Medicine (DOM). R.B.A. is supported by NIH/National Institute of General Medical Sciences PharmGKB resource, R24GM61374, as well as LM05652 and GM102365. Additional support is from the Stanford NIH/National Center for Research Resources CTSA award number UL1 RR025744.

Patient data extracted and de-identified by Tanya Podchiyska of the STRIDE (Stanford Translational Research Integrated Database Environment) project, a research and development project at Stanford University to create a standards-based informatics platform supporting clinical and translational research. The STRIDE project described was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through grant UL1 RR025744. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Services, H. Health information technology: standards, implementation specifications, and certification criteria for electronic health record technology, 2014 edition; revisions to the permanent certification program for health information technology. Final rule. *Fed. Regist.* **77,** 54163–292 (2012).
2. Kaushal, R., Shojania, K. G. & Bates, D. W. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. *Arch. Intern. Med.* **163,** 1409–1416 (2003).
3. Overhage, J. & Tierney, W. A randomized trial of "corollary orders" to prevent errors of omission. *J. Am. Med. Informatics Assoc.* **4,** 364–75 (1997).
4. Bates, D. & Kuperman, G. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. *J. Am. Med. Informatics Assoc.* **10,** 523–530 (2003).
5. De Lissovoy, G. Big data meets the electronic medical record: a commentary on "identifying patients at increased risk for unplanned readmission". *Med. Care* **51,** 759–60 (2013).
6. Moore, K. D., Eyestone, K. & Coddington, D. C. The big deal about big data. *Healthc. Financ. Manage.* **67,** 60–6, 68 (2013).
7. Doddi, S., Marathe, a, Ravi, S. S. & Torney, D. C. Discovery of association rules in medical data. *Med. Inform. Internet Med.* **26,** 25–33 (2001).
8. Klann, J., Schadow, G. & Downs, S. M. A method to compute treatment suggestions from local order entry data. *AMIA Annu. Symp. Proc.* **2010,** 387–91 (2010).
9. Klann, J., Schadow, G. & McCoy, J. M. A recommendation algorithm for automating corollary order generation. *AMIA Annu. Symp. Proc.* **2009,** 333–7 (2009).
10. Wright, A. & Sittig, D. F. Automated development of order sets and corollary orders by data mining in an ambulatory computerized physician order entry system. *AMIA Annu. Symp. Proc.* **2006,** 819–823 (2006).
11. Chen, J. R. A. Mining for Clinical Expertise in (Undocumented) Order Sets to Power an Order Suggestion System. *Proc. 2013 AMIA Summit Clin. Res. Informatics* (2013).
12. Linden, G., Smith, B. & York, J. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Comput.* **7,** 76–80 (2003).
13. Lowe, H. J., Ferris, T. a, Hernandez, P. M. & Weber, S. C. STRIDE--An integrated standards-based translational research informatics platform. *AMIA Annu. Symp. Proc.* **2009,** 391–5 (2009).
14. Wright, A. & Bates, D. W. Distribution of Problems, Medications and Lab Results in Electronic Health Records: The Pareto Principle at Work. *Appl. Clin. Inform.* **1,** 32–37 (2010).
15. Manning, C. D. & Schütze, H. *Foundations of Statistical Natural Language Processing*. *Comput. Linguist.* **26,** 277–279 (MIT Press, 1999).
16. Shani, G. & Gunawardana, A. Evaluating Recommendation Systems. *Recomm. Syst. Handb.* **12,** 1–41 (2011).
17. Lemeshow, S. & Le Gall, J. R. Modeling the severity of illness of ICU patients. A systems update. *JAMA* **272,** 1049–55 (1994).
18. Renaud, B. *et al.* Risk stratification of early admission to the intensive care unit of patients with no major criteria of severe community-acquired pneumonia: development of an international prediction rule. *Crit. Care* **13,** R54 (2009).

Considerations for Using Research Data to Verify Clinical Data Accuracy

Daniel Fort, MPH[1], Chunhua Weng, PhD[1], Suzanne Bakken, RN, PhD[1,2], Adam B. Wilcox, PhD[3]

[1]Department of Biomedical Informatics, [2]School of Nursing, Columbia University, New York City; [3]Intermountain Healthcare, Salt Lake City, UT

Abstract

Collected to support clinical decisions and processes, clinical data may be subject to validity issues when used for research. The objective of this study is to examine methods and issues in summarizing and evaluating the accuracy of clinical data as compared to primary research data. We hypothesized that research survey data on a patient cohort could serve as a reference standard for uncovering potential biases in clinical data. We compared the summary statistics between clinical and research datasets. Seven clinical variables, i.e., height, weight, gender, ethnicity, systolic and diastolic blood pressure, and diabetes status, were included in the study. Our results show that the clinical data and research data had similar summary statistical profiles, but there are detectable differences in definitions and measurements for individual variables such as height, diastolic blood pressure, and diabetes status. We discuss the implications of these results and confirm the important considerations for using research data to verify clinical data accuracy.

Introduction

Computational reuse of clinical data from the electronic health record (EHR) has been frequently recommended for improving efficiency and reducing cost for comparative effectiveness research[1]. This goal faces significant barriers because clinical data are collected to aid individual clinicians in diagnosis, treatment, and monitoring of health-related conditions rather than for research uses[2]. A risk to reuse is potential hidden biases in clinical data. While specific studies have demonstrated positive value in clinical data research, there are concerns about whether they are generally usable. An opaque data capture processes and idiosyncratic documentation behaviors of clinicians from multiple disciplines may lead to data biases. A difference in the population who seek medical care versus the general residential population may introduce a selection bias when clinical data are used to estimate population statistics.

Comparison of EHR data with a gold standard is by far the most frequently used method for assessing accuracy[3]. Recent efforts have taken a more implicit approach to validating clinical data in the form of study result replication. Groups such as HMORN, OMOP, and DARTNet assessed the accuracy of clinical data by comparing research results derived from clinical data with those derived from randomized controlled trials[4-6]. This reflects a focus on making a new system work, rather than a lack of recognition of a potential problem.

The Washington Heights/Inwood Informatics Infrastructure for Community-Centered Comparative Effectiveness Research (WICER) Project (http://www.wicer.org) has been conducting community-based research and collecting patient self-reported health information. We assume research data are of better quality than clinical data given their rigorous data collection processes. For patients with information in both the survey and electronic health records, an analysis of the differences between data collected through survey and data collected in clinical settings may help us understand the potential biases in clinical data. This study compares WICER Community Survey results to data for the same variables from the same people collected within our EHR as well as attempts to replicate the WICER research sample using only clinical data. We discuss the implications of these results and three potential categories of accuracy of clinical data.

Methods

Our conceptual framework for using research data to verify clinical data includes four consecutive steps: (1) cohort selection; (2) variable selection; (3) data point selection; and (4) measurement selection.

Step 1: Cohort Selection

We selected the patients who had data in both data sources: the WICER community population health survey and our institutional clinical data warehouse. The WICER Community Survey collected data from residents in Washington Heights, an area of New York with a population of approximately 300,000 people, through cluster and

snowball sampling methodologies. Surveys were administered to individuals over the age of 18 who spoke either English or Spanish. Survey data was collected and processed from March 2012 through September 2013. A total of 5,269 individuals took the WICER Community Survey in either the Household or Clinic setting.

The Columbia University Medical Center's Clinical Data Warehouse (CDW) integrates patient information collected from assorted EHR systems for about 4 million patients for more than 20 years. The initial effort to replicate the WICER research sample restricted the CDW to adult patients who had an address within one of the five same zip codes and one recorded visit during the WICER data collection time period, resulting in a cohort of 78,418 patients.

The WICER data set includes a higher proportion of women and Hispanic individuals than either the CDW sample or what was expected based on census data for the same area codes. New clinical data samples were created to match the proportion of women and Hispanic ethnicity as found in the WICER data set, as well as new samples for both which match the census distributions for age and gender. A total of 1,279 individuals were identified from the intersection of the two datasets to compare clinical data in CDW and research data in WICER without a sampling bias.

Concept	Definition
N	Number of individuals in sample
Age	Average age of individuals in sample
Proportion Female	Proportion of sample labeled female
Proportion Hispanic	Proportion of sample labeled Hispanic
Weight kg	Average weight of individuals in sample
Height cm	Median height of individuals in sample
BMI	Average BMI of individuals in sample, computed from individual weight and height
Prevalence of Smoking	Proportion of sample labeled positive for smoking
Prevalence of Smoking with labeled status	Proportion labeled positive for smoking out of individuals with labeled smoking status
Systolic	Average systolic blood pressure of individuals in sample
Diastolic	Average diastolic blood pressure of individuals in sample
Prevalence of Diabetes, Strict Criteria	Proportion of sample with positive diabetes status.
Prevalence of Diabetes among labeled status, Strict Criteria	Proportion with positive diabetes status out of individuals with recorded ICD-9 codes and test values

Table 1: Concepts and Definitions for sample summary

Step 2: Variable Selection

Because the WICER study included variables related to hypertension, the American Heart Association (AHA) / American College of Cardiology (ACC) original guidelines for cardiac risk were chosen to guide the variable selection process[7]. The content overlap between the wide range of information collected for the WICER Community Survey and that available in the CDW is limited to some basic demographic and baseline health information. Of the factors in the AHA/ACC Guidelines, Age, Race, Ethnicity, Gender, the components of BMI (height and weight), Smoking Status, Blood Pressure (systolic and diastolic) were available as structured data in both data sources. See Table 1 for concept definitions.

A simple clinical phenotyping method, consistent with the eMERGE diabetes phenotype[8] but excluding medication orders, was developed for type 2 diabetes in the CDW using ICD-9 Codes, HbA1c test values, and glucose test values. Using the strictest criteria, a patient will only be identified as having diabetes if there are at least two ICD-9 codes for diabetes, at least one HbA1c test value >6.5, or at least two high glucose test values. A glucose test value is coded as high if it is >126 for a fasting glucose test or >200 otherwise. Effectiveness of labeling of each of these components was also explored.

Step 3: Data Point Selection

Each clinical variable could have many data points from multiple points of measurement across time, which necessitated careful data point selection to ensure that summary data points were both representative of all data points and comparable across data sources without introducing data sampling biases. This includes an issue of temporal bias, where some data variables, such as weight, might naturally be expected to change over time. To make a comparable cross-section to the Survey dataset and to ensure the resulting data reflects not only the same sample but also the same sample at the same time, we selected only data points recorded during the 18-month WICER study period from the CDW. In this way, assuming the survey participants are measured at random throughout an 18-month period, so too are the clinical data population.

In the matched sample we had an opportunity to more finely tune the data comparison. The most direct approach is to simply select the clinical data point closest in time to the survey measurement of any given participant.

Alternatives include the closest prior or subsequent data paint as well as using a single randomly selected point rather than the average of all clinical data points. While alternate data point selection options were explored, to best keep the results comparable the reported values for the matched sample were derived in the same fashion as for the sample at large.

Step 4: Data Measure Selection for Comparing the Two Data Sets

With representative patient sample, meaningful variables, and representative data points, the next important step for designing an unbiased verification study was to select a meaningful data measure, which seems to be the most subjective step without standard guidance. For this step, we considered two measures: (a) population-level average summary statistics; and (b) patient-level average summary statistics.

Option (a): Population-Level Average summary statistics

Multiple data values available during the study period were averaged in order to minimize any temporal effects while also allowing the use of the most number of patients. Continuous variables within each set were averaged, with one exception, and compared via t-test. The median BMI value was used for comparison as the mean summary value for the calculation of BMI is more susceptible to outliers. Choice of other "best matching" clinical data values, such as the closest prior and subsequent values in time as well as simple random choice, were also explored.

Proportions of interest, which include % female, % smoking, and % Hispanic, for the categorical variables were reported and compared with chi-square test. For some proportions there is a possibility that negative or healthy status might not be recorded and would therefore be accurately represented by missing data. Therefore for smoking and diabetes there is a second value reported: the proportion of labeled status, which excludes any patient with missing data rather than assume missing data denotes known negative status.

For the purpose of primary analysis, only the strictest, ALL criteria for diabetes diagnosis are reported, as consistent with the eMERGE criteria. However, each component of the diabetes diagnosis was examined for sensitivity, specificity, and positive predictive against the patient's self-reported diabetes status. All summary and statistical comparisons were performed in Python, using the SciPy scientific computing package for statistical comparisons.

Option (b): Patient-level Average Summary Statistics

When there is sufficient clinical data, it is possible to create a distribution of expected values for a given patient and compare the survey value to that distribution. At its simplest, the comparison is simply whether the survey value is within one standard deviation of the mean of the available clinical values. This process was performed for patients with at least five data points for the same variable recorded during the study period.

Results

Summary values for the WICER Survey population, the raw clinical sample, the resampled clinical data targeted to match the survey proportion of women and

	Survey	Census-weighted Survey		Clinical Raw	Clinical Resampled	Census-weighted Clinical
N	4069			78418	56694	
Age	50.1	44.6		47.6	47.0	44.1
Proportion Female	0.708	0.528		0.619	0.714	0.528
Proportion Hispanic	0.955	0.951		0.496	0.604	0.501
Weight kg	75.4	77.0		75.7	74.8	78.2
Height cm	161.2	163.7		160.3	159.1	162.7
BMI	28.2	27.7		28.1	28.3	28.1
Prevalence of Smoking	0.058	0.064		0.089	0.078	0.101
Prevalence of Smoking with labeled status	0.060	0.066		0.122	0.103	0.138
Systolic	127.7	125.5		127.2	126.5	126.8
Diastolic	81.0	80.7		73.1	72.7	73.4
Prevalence of Diabetes, Strict Criteria	0.159	0.122		0.038	0.040	0.032
Prevalence of Diabetes among labeled status, Strict Criteria	0.162	0.124		0.284	0.286	0.313

Table 2: Summary Results from alternate sampling methods

hispanic participants, and census distribution weighted samples are presented in **Table 2**. Analysis was performed across all samples with no significant variation in results. The original, total clinical dataset was chosen for representative purposes because it is the only clinical sample to contain all members of the matched set.

Following the population summary approach, values and statistics for each data point are presented in **Table 3**. The Survey dataset tends to be slightly older and contain more women. Survey participants were almost entirely identifying as Hispanic. Sixteen percent of the survey participants self-identified as having diabetes. Measuring the Matched dataset via clinical data and primary survey collection processes broadly records the same values. There are statistically significant measurement discrepancies in Hispanic ethnicity labeling, height measurement, diastolic blood pressure, and diabetes status determination. Where the Clinical and Survey datasets differ, in age, proportion of women, and prevalence of smoking, are evidence of statistically significant differences in sample composition. In exploring patient-level summary statistics, the number of patients with sufficient data to construct a distribution of expected blood pressures was 866. Of these, 491(57%) and 479(55%) had a survey systolic or diastolic blood pressure, respectively, greater than one standard deviation away from their clinical mean. **Table 4** shows an example result of alternate data point selections in Systolic BP. While values are statistically significantly different from one another in this and other examples, they would not change the conclusions drawn from **Table 3**.

	Clinical	Matched Clinical	Matched Survey	Survey		p-value: Matched vs. Matched	p-value: Clinical vs. Survey
N	78,418	1,279		5,269			
Age	47.55	52.33	51.12	50.12		0.072	**p << .0001**
Proportion Female	0.62	0.79	0.78	0.71		0.963	**p << .0001**
Proportion Hispanic	0.50	0.56	0.94	0.96		**p << .0001**	**p << .0001**
Weight kg	75.69	77.16	76.99	75.42		0.851	0.851
Height cm	160.34	158.23	161.31	161.25		**p << .0001**	**p << .0001**
BMI	28.10	29.70	28.90	28.20		0.207	0.207
Prevalence of Smoking	0.09	0.08	0.08	0.06		0.944	**p << .0001**
Prevalence of Smoking with labeled status	0.12	0.09	0.08	0.06		0.283	**p << .0001**
Systolic	127.23	128.48	127.50	127.68		0.204	0.164
Diastolic	73.07	74.34	79.24	80.95		**p << .0001**	**p << .0001**
Prevalence of Diabetes, Strict Criteria	0.04	0.09	0.22	0.16		**p << .0001**	**p << .0001**
Prevalence of Diabetes among labeled status, Strict Criteria	0.28	0.35	0.23	0.16		0.001	**p << .0001**

Discrepancy in Measurement Discrepancy in Selection

Table 3: Clinical, Survey, and Matched Set data comparison. Bonferroni-corrected p-value = 1e-4

Systolic BP	Survey	Closest Prior	Closest Subsequent	Random Point	Mean
N	1290	1107	962	1185	1185
Mean	127.8	127.9	130.3	129.3	128.5

Table 4: Systolic blood pressure summary values and patient cohort size for various data point selection methodologies

Value	ALL	ANY	≥1 ICD-9	≥2 ICD-9	HIGH HBA1C	HIGH GLUCOSE, EVER	HIGH GLUCOSE, RECENT
Sensitivity	0.33	0.81	0.90	0.84	0.48	0.72	0.52
Specificity	0.98	0.35	0.88	0.93	0.96	0.53	0.74
F-measure	0.49	0.49	0.89	0.88	0.64	0.61	0.61
Positive Predictive Value	0.82	0.27	0.68	0.78	0.79	0.31	0.37

Table 5: Sensitivity, Specificity, F-measure, and Positive Predictive Value of components of a diabetes diagnosis

The sensitivity, specificity, and positive predictive value of various strategies to identify diabetes status using clinical data are presented in **Table 5**. In this simple phenotype, ALL is the intersection of three criteria and ANY is the union. The three criteria are having at least two ICD-9 codes for diabetes, one high HbA1c value, and at least two high glucose values. The rationale for requiring two of some categories is to restrict potentially spurious results. In the case of diagnostic codes, for example, a diabetes ICD-9 code might be recorded for a negative diabetes

evaluation. The removal of these restrictions was also considered. The ALL criteria have the highest positive predictive value, but the lowest sensitivity. Both the ICD-9 and HbA1c-based criteria have high specificities and the ICD-9 based criteria alone have the highest F-measure for sensitivity and specificity. Proportions of patients retrieved under each qualifying criteria are consistent with published results[9].

Discussion

Our study shows discrepancies between clinical and research data, both in sampling and measurement. Clinical measurement of some data, such as gender and BMI, accurately reproduces the research measurement and others, such as diabetes, do not. While raw results may be interesting, because of the limits of overlapping data between sets and the comparisons which could be made, the raw results may have little value outside of this case study. If these discrepancies can be considered as representative of classes of clinical data, we can abstract some idea of generalizable accuracy of clinical data as compared to primary research data. We introduce three categories of accuracy.

The first category is "completely accurate" information, such as sex, birthdate, and therefore age. These data might be considered Personally Identifiable Information (PII), or information that on its own could be used to identify an individual. This classification suggests that address, social security number, and phone number would also be accurate between datasets. While there will be instances of coding error, misreporting, or other errors, by and large these data are consistent across datasets. It should be noted that birthdate was one of the criteria by which individuals were identified for the Matched, and therefore errors in the recording of birthdate would be excluded from this analysis. Also, while PII should be accurate across datasets, this does not suggest that all demographic information, such as ethnicity, will be accurate.

The second category is 'simple measurement' information, which is the result of a clear concept or measurement process. Height, weight, systolic and diastolic blood pressure, smoking status, and ethnicity are included in this category. Here, the simplicity of the measurement or concept leads to agreement in the value between sources, and differences in the value are the result of a difference in either the concept definition or the measurement process. For example, measured heights in the Matched group differ by approximately 2.5cm or 1in, suggesting that the concept and measurement of height in the Survey sample includes shoes. Likewise, diastolic blood pressure is consistently measured 5 points higher in the Survey sample, suggesting a difference in measurement. Ethnicity, which is self-reported in the survey, is labeled by hospital staff during admission to the hospital, resulting in approximately one third of Hispanic individuals being labeled as 'Unknown' ethnicity in the Clinical sample.

The final category of accuracy is 'inferred' information, where a complex concept, such as diabetes, is inferred from multiple variables. When compared with self-reported Survey values, no single prediction or combination of variables can be considered accurate for an entire cohort. However, some results may be useful enough for a specific purpose. For example, requiring ALL criteria has a high positive predictive value and may provide a high level of accuracy within a given cohort. Conversely, using just HbA1c measurements has a high sensitivity and may be most valuable when a larger quantity of data is required for statistical power.

At least in this case study, discrepancies in the 'simple measurement' category are stable across multiple sampling methodologies. Discrepancies are also stable when samples are broken down into categories such as age by decade, obesity classification, and hypertension risk category. This stability is what would be expected if the discrepancies were the result of simple measurement error and would suggest these discrepancies represent systematic bias in the clinical data. It is possible that reported discrepancies are the result of data retrieval and processing. However, the presence of pairs of measurements such as weight/height and systolic/diastolic blood pressure, retrieved and processed in an identical manner, where one is accurate and one not, suggests the discrepancies are truly present in at data source. Due to the limitations of this case study, it is unclear how generalizable this finding may be.

The choice of exact data points may also influence study results, so care must be taken in accurately summarizing patient data. In this study, the biggest apparent difference was between closest prior and subsequent data points. The reason may be that closest prior data point represents the end of a series of blood pressures which began with a hospitalization and is, therefore, the nearest to "normal". The closest subsequent data point, however, would represent the initial data collection of a hospitalization and would likely reflect a health crisis. Furthermore, defining allowable data points in time restricts the number of patients, who qualify for comparison. Using the average value for each patient smoothens out these temporal effects and allows the use of the maximum number of patients for comparison.

Recommendations

When a research cohort is defined as having clinical data, that clinical data may be a usable substitute for primarily acquired research data, depending on the needs of the research. PII should have a high degree of accuracy and aspects of the patient record, which are conceptually simple or have a clear measurement process, may be accurate or include relatively small discrepancies. More complex concepts, such as a diabetes phenotype, are not accurate for summary purposes but components may be useful depending on the exact nature of the requirement. However, to avoid the discrepancies due to clinical sampling, the research cohort must be defined as already having clinical data. Results of this study demonstrate a significant difference in sampling processes between clinical data and research survey cohorts. Clinical data used as a convenience sample to substitute for primary research data will not accurately describe the target population. Discrepancies in the simple measurement category may be due to differences between either the concept definition or measurement processes. If a dictionary of concept definitions or measurement procedures was provided as either a standalone document or as metadata tied to each value, such as whether a measurement of height requires shoes to be taken off, then the comparability of specific variables might be predictable. Additionally, while aspects of clinical data collection are not in the researcher's control, the exact choice of data value for research may be. Different choices, such as average per patient or nearest in time, can result in statistically significant differences in values.

Limitations

This study is limited in scope and setting. First, the overlap between the population survey and clinical data was limited to a small set of variables. Second, the population survey targeted a largely Hispanic, urban population and the institution is a large academic medical center. These findings may not be generalizable to other institutions and populations. This work should be replicated in other settings.

Conclusions

We compared research population survey results to clinical data for the same target population to verify the accuracy of clinical data elements. Clinical data elements may be classified into three categories of accuracy: completely accurate, simple measurement, and inferred information, depending in part on the complexity of the concept being measured and the process of that measurement. Additionally, we report recommendations and considerations for using clinical data for cohort selection and research.

Acknowledgments

The authors are supported by grant R01HS019853 (PI: Bakken) from the Agency for Healthcare Research and Quality, grants 5T15LM007079 (PI: Hripcsak) and R01LM009886 (PI: Weng) from the National Library of Medicine, and grant UL1 TR000040 (PI: Ginsberg) from the National Center for Advancing Translational Sciences.

References

1. *A First Look at the Volume and Cost of Comparative Effectiveness Research in the United States*, 2009, Academy Health.
2. Hersh, W.R., M.G. Weiner, P.J. Embi, J.R. Logan, P.R. Payne, E.V. Bernstam, H.P. Lehmann, G. Hripcsak, T.H. Hartzog, J.J. Cimino, and J.H. Saltz, *Caveats for the use of operational electronic health record data in comparative effectiveness research*. Med Care, 2013. **51**(8 Suppl 3): p. S30-7.
3. Weiskopf, N.G. and C. Weng, *Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research*. J Am Med Inform Assoc, 2013. **20**(1): p. 144-51.
4. *OMOP Design and Validation*. 2013; Available from: http://omop.fnih.org.
5. Tannen, R.L., M.G. Weiner, and D. Xie, *Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings*. BMJ, 2009. **338**: p. b81.
6. Libby, A.M., W. Pace, C. Bryan, H.O. Anderson, S.L. Ellis, R.R. Allen, E. Brandt, A.G. Huebschmann, D. West, and R.J. Valuck, *Comparative effectiveness research in DARTNet primary care practices: point of care data collection on hypoglycemia and over-the-counter and herbal use among patients diagnosed with diabetes*. Med Care, 2010. **48**(6 Suppl): p. S39-44.
7. Grundy, S.M., R. Pasternak, P. Greenland, S. Smith, Jr., and V. Fuster, *AHA/ACC scientific statement: Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology*. J Am Coll Cardiol, 1999. **34**(4): p. 1348-59.

8. Pacheco, J.T., W. *Type 2 Diabetes Mellitus*. 2012; Available from: http://phenotype.mc.vanderbilt.edu/phenotype/type-2-diabetes-mellitus.
9. Richesson, R.L., S.A. Rusincovitch, D. Wixted, B.C. Batch, M.N. Feinglos, M.L. Miranda, W.E. Hammond, R.M. Califf, and S.E. Spratt, *A comparison of phenotype definitions for diabetes mellitus*. J Am Med Inform Assoc, 2013. **20**(e2): p. e319-26.

How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?

Preethi Raghavan, MS, James L. Chen, MD, Eric Fosler-Lussier, PhD, Albert M. Lai, PhD
The Ohio State University, Columbus, OH

Abstract

Electronic health records capture patient information using structured controlled vocabularies and unstructured narrative text. While structured data typically encodes lab values, encounters and medication lists, unstructured data captures the physician's interpretation of the patient's condition, prognosis, and response to therapeutic intervention. In this paper, we demonstrate that information extraction from unstructured clinical narratives is essential to most clinical applications. We perform an empirical study to validate the argument and show that structured data alone is insufficient in resolving eligibility criteria for recruiting patients onto clinical trials for chronic lymphocytic leukemia (CLL) and prostate cancer. Unstructured data is essential to solving 59% of the CLL trial criteria and 77% of the prostate cancer trial criteria. More specifically, for resolving eligibility criteria with temporal constraints, we show the need for temporal reasoning and information integration with medical events within and across unstructured clinical narratives and structured data.

Introduction

The electronic health record (EHR) is a powerful repository of patient information that can be leveraged to build applications that benefit the clinical community such as clinical trial recruitment. Understanding and extracting information from EHRs enables reasoning with clinical variables and supports decision making.[1] EHRs record patient information both as data coded in structured format, as well as in the form of free text clinical narratives. Structured data typically contains demographics, patient birth and death information, lab values, encounters, and at times procedures and diagnosis lists. Unstructured data includes free text clinical narratives that correspond to different encounters generated at various points of time, including admission notes, history and physical reports, discharge summaries, radiology reports, and pathology reports.

Clinical trial recruitment may be semi-automated through information extraction from the EHR. Clinical trials have eligibility (inclusion and exclusion) criteria that describe characteristics and constraints that help determine if a patient qualifies for a trial. Typically, clinicians and trial recruitment coordinators identify potential clinical trial patients from characteristics described in their medical history and match them against the eligibility criteria for individual trials. This standard model of clinical trial enrollment is rife with errors. If the clinical staff is unfamiliar with a particular trial or if there are competing trials, an eligible patient may be overlooked. On the opposite extreme, the clinical trials staff may be asked to evaluate patients who are clearly not candidates.

This information mismatch has the potential to be streamlined. Generating automated queries corresponding to eligibility criteria and querying patient records from the EHR in order to identify qualifying patients provides an efficient and agnostic approach to clinical trials recruitment..The pertinent question then is whether structured data, being easier to automatically process and understand, has sufficient information to resolve these eligibility criteria, or if there is a need to extract and reason with medical concepts in unstructured clinical narratives. Researchers have often emphasized the importance of using clinical narratives for clinical decision support,[1] information retrieval,[2] question answering[1] and automated clinical trial recruitment.[4] Unstructured data in clinical narratives captures important decisions and relationships between medical concepts including causal (symptom caused disease), consequential (why a drug or treatment was administered) and temporal (symptom before disease/ treatment). Furthermore, Rosenbloom et al.[5] suggest that clinical notes containing naturalistic prose have been more accurate and reliable for identifying patients with given diseases, and more understandable to healthcare providers reviewing patient records. However, to the best of our knowledge, there are no prior empirical studies that evaluate the usefulness of structured vs. unstructured data considering their advantages and limitations for a clinical task.

In this paper, we study two datasets of structured and unstructured data with patients suffering from chronic lymphatic leukemia (CLL) and prostate cancer obtained from The Ohio State University Wexner Medical Center. Given a set of eligibility criteria from corresponding clinical trials, we evaluate the number of criteria that can be resolved using information from just the structured data and the number of criteria that require information extraction from and reasoning with unstructured clinical narratives and data. There are three main contributions of this work: 1) Empirical evaluation of the commonly assumed hypothesis that unstructured clinical text processing is required and that structured data alone is insufficient to accurately resolve eligibility criteria with the help of a clinical trial use case; 2) Demonstration of the need for cross-narrative temporal reasoning in solving certain

temporal eligibility criteria; 3) Demonstration of the need for information fusion across structured and unstructured data in solving certain temporal eligibility criteria.

Related Work

The recent decade has seen considerable research in the natural language processing (NLP) of unstructured clinical text.[3,6-8] Fushman et al.[1] discuss how successful processing of clinical narratives is the key to overall success of automated clinical decision support systems. They stress the importance of medical concepts with the help of named entity recognition and learning relations between those named entities are important for better understanding clinical narrative text. Wang et al.[7] propose a framework for automated pharmacovigilance by applying NLP and association statistics on comprehensive unstructured clinical data from the EHR. They argue that previous algorithms have focused on coded and structured data, and therefore miss important clinical data relevant to this task. Medical NLP systems like Mayo's cTakes,[8] and MedLEE[7] have components specifically trained or designed for information extraction from clinical text.

There has been some work on modeling temporal knowledge in eligibility criteria to help effective clinical text processing. [9-10] Ross et al. [10] observe that temporal features were present in 40% of clinical trial criteria analyzed as part of their study, where the type of temporal expression in the criteria ranged from well-specified to loosely-specified. Similarly, there have been considerable efforts, including rule-based algorithms, temporal annotation of clinical corpora, and machine learning methods, towards learning temporal relations and generating timelines of medical events from unstructured clinical text.[11-13] Zhou et al.[11] extract temporal relations between medical events in discharge summaries. The CLEF project[12] uses a pairwise supervised classification approach to learn temporal relations between medical events within the same narrative. While temporal information has been studied in the intra-document context, there is not much prior work in cross-narrative temporal relation learning and information fusion. Carlo et al.[14] attempt to align medical problems in structured and unstructured EHR data using UMLS by studying the information overlap between structured ICD-9 diagnoses and unstructured discharge summaries. They conclude that this is a non-trivial task with the need for better methods to detect correlating structured and unstructured data before aligning them. Köpcke et al.[15] compare the eligibility criteria defined in trial protocols with patient data contained in the EHR in multi-site trials to determine the extent of available data compared with the eligibility criteria of randomly selected clinical trials. However, their study is restricted to structured data in the EHR.

In spite of the large body of recent work in processing structured and unstructured clinical narratives for temporal reasoning, and other NLP tasks, there are no prior studies that empirically evaluate the usefulness of structured vs. unstructured data for a clinical task. We perform an empirical analysis of CLL and prostate cancer patient records and evaluate the performance of structured and unstructured data in resolving clinical trial eligibility criteria. We specifically focus on criteria with temporal constraints and illustrate the need for unstructured clinical narrative analysis including cross-narrative temporal reasoning and information fusion.

Patient Records and Clinical Trial Eligibility Criteria - Data Description

The EHR data used in this study consists of medical records for 2060 CLL patients and 1808 prostate cancer patients. The CLL dataset contains 95 different types of unstructured reports including discharge summaries, history and physical reports, specialty reports such as wound care, operative notes, OB/GYN and psych evaluations, social work assessment, referral letters and progress notes. It also consists of radiology reports, pathology reports and cardiology reports. The total number of unstructured clinical narratives in the CLL dataset is 100704. The structured data consists of lab reports, procedures list, diagnoses list and encounters list.

The prostate cancer dataset consists of 2652 oncology reports, 1582 pathology reports, 6606 radiology reports as part of unstructured data. The structured data in this dataset includes a discharge medications list (30178 medications), laboratory values (939 values), and a medications list (141932 medications).

The clinical trials dataset consists of a set of top 100 clinical trials each, as defined by clinicaltrials.gov, for both CLL and prostate cancer.

Methodology

Medical concept extraction - We annotated the clinical trial criteria datasets with medical concepts, concept unique identifiers (CUIs) and semantic types using MetaMap.[14] We then extracted criteria containing the following semantic types: Disease or Syndrome, Laboratory or Test Result, Procedure, Sign or Symptom, and Pharmacological Substance. The criteria containing the Temporal Concept semantic type were labeled as temporal eligibility criteria. Similarly, we also annotated both patient datasets with medical concepts and the semantic types mentioned previously.

Matching medical concepts across clinical trials and patient datasets - In order to evaluate the degree of overlap between the clinical trials dataset and structured and unstructured data in the medical records dataset, we compute

the *Match* between medical concepts across these datasets. The match functions are computed across the datasets as follows. 1) UMLS CUI Match where an exact CUI match is computed and 2) Phrase Match where we compute a match between medical concepts (textual fragment identified as the medical concept). Thus we have,

- *Match*(CUI in the trial dataset, CUI in structured data)
- *Match*(CUI in the trial dataset, CUI in unstructured data)
- *Match*(Phrase in the trial dataset, medical concept in the structured data)
- *Match*(Phrase in the trial dataset, medical concept in the unstructured data)

These match functions are computed for two levels of analysis - (1) medical concept-level, where we compare all the medical concepts in the trials dataset against the structured and unstructured data, and (2) eligibility criteria level, where we compare all the medical concepts in each criterion against the structured and unstructured data.

The medical concept-level match helps analyze the number and type of medical concepts typically found in the structured and unstructured datasets when solving clinical trial eligibility criteria. As shown in the algorithm below, we compute the *match* between all medical concepts in the clinical trials dataset and the structured data. If there are no matching concepts found in the structured data, we then compute a *match* with the unstructured data.

1. Calculate
   a. *Match*(CUI in the trial dataset, CUI in the structured data)
   b. *Match*(Phrase in the trial dataset, medical concept in the structured data)
2. If there are no match results from step 1, then calculate
   a. *Match*(CUI in the trial dataset, CUI in the unstructured data)
   b. *Match*(Phrase in the trial dataset, medical concept in the unstructured data)

The eligibility criteria-level match helps us analyze the number of criteria that can be solved by structured data, unstructured data or both. In order to evaluate the need for temporal reasoning and information fusion and constrain the number of eligibility criteria, we restricted the eligibility criteria-level analysis to criteria with temporal constraints. We compare each eligibility criterion against both structured data and unstructured data to determine if the concepts in the criterion require only structured data, only unstructured data or both datasets together for resolution, as shown in the algorithm below.

1. For all temporal eligibility criteria,
   a. For all medical concepts (from 1 to *n*) in the criterion
      i. $Match_1$(CUI in the criterion, CUI in the structured data) ∧ … ∧ ($Match_n$(CUI in the criterion, CUI in the structured data)
      ii. $Match_1$(Phrase in the criterion, Phrase in the structured data) ∧ … ∧ ($Match_n$(Phrase in the criterion, Phrase in the structured data)
2. If i OR ii returns *true,* then the criterion can be resolved by the structured data
3. Repeat step 1. by replacing "structured data" with "unstructured data"
   a. If step i OR ii returns *true,*
      i. the criterion can be resolved by the unstructured data
      ii. else the criterion can be cannot be resolved by a concept match across unstructured data
4. If in step 2, we get *true* for "structured" as well as "unstructured data",
   a. the criterion can be solved using *either* the structured or unstructured data.

The algorithm first compares all medical concepts in the eligibility criterion against all medical concepts in the structured data. If all the concepts in the criterion are found in the structured data, we conclude that the criterion may be resolved using the structured data. We then do a similar comparison for unstructured data and if all concepts in the criterion are found in the unstructured data, we conclude that the criterion may be resolved using the unstructured data.

Information fusion - In the case where all the concepts in the criterion are found in both the structured as well as the unstructured data, we conclude that the criterion can be solved using *either* the structured or the unstructured data. However, the criterion may also require both structured as well as unstructured data for resolution. Taking this into consideration, we define information fusion as follows.

Given medical concepts $\{m_1, \ldots, m_n\}$ in a clinical trial criterion, if S_k is a set of k concepts that match the structured data and U_j is a set of j concepts that match the unstructured data, where $k, j>0$ and $k, j<n$. Now there are two possibilities.

1. $L = S_k \cap U_j$ is not empty. Here, L concepts match both structured and unstructured data.
2. $L = S_k \cap U_j$ is empty. Here, L concepts match the structured data and the remainder j concepts match the unstructured data. So S_k and U_j are disjoint.

Temporal reasoning in unstructured data - For subset of criteria that require unstructured data for resolution, we further analyze the temporal constraints in the criteria and attempt to answer the following questions. How many temporal constraints can be solved using coarse temporal reasoning within each clinical narrative? How many temporal constraints require more granular temporal ordering within each clinical narrative? How many temporal constraints require cross-narrative temporal reasoning?

In order to answer these questions, we run a CRF-based time-bin tagger[17] and learn to associate the medical events within each narrative with one of the coarse time-bins: "*way before admission, before admission, admission, after admission, discharge*". The time-bin tagger was trained on different patient records not part of this dataset. We also perform fine-grained temporally ordering by learning to rank medical concepts within a clinical narrative by their order of occurrence.[18] This gives us both a coarse ordering and a fine-grained ordering of medical concepts within each clinical narrative. These intra-narrative temporal orderings are then combined with the admission and discharge dates across narratives to generate a cross-document partially ordered timeline of medical concepts for each patient.

Results

The methodology is empirically evaluated by calculating the extent of *match* between the eligibility criteria dataset and the structured and unstructured datasets. The medical concept-level match results between the trials datasets, consisting of all eligibility criteria, and the structured and unstructured data are shown in Table 1. The CLL trials dataset has 2167 medical concepts and the prostate cancer dataset has 1019 medical concepts.

The CLL trials have a total of 1720 eligibility criteria, while the prostate cancer trials have 1325 eligibility criteria, containing diseases, procedures, tests, symptoms and medications. We observe that more than half of the medical concepts in the CLL and prostate patient data were only found in the unstructured data. The most frequent medical concept semantic types found in the unstructured datasets include Finding, Sign or Symptom, Disease or Syndrome, whereas the most frequent medical concept semantic type in the structured data includes Laboratory Test or Procedure, Pharmacological Substance and Disease or Syndrome. If the structured data has diagnoses and encounters lists, there tend to be overlapping Disease or Syndrome type concepts across the structured data and unstructured clinical narratives.

	CLL		Prostate Cancer	
	CUI	Medical Concept	CUI	Medical Concept
Structured Data Match	23%	29%	11%	19%
Unstructured Data Match	61%	68%	48%	57%

Table 1: Medical Concept-level Analysis on CLL and Prostate Cancer Trials and Patient Records

354 of the eligibility criteria in the CLL trials and 297 of the eligibility criteria in the prostate cancer trials have temporal constraints. Table 2 shows results from matching temporal clinical trial eligibility criteria against structured and unstructured data. In both patient datasets, matching the textual fragment identified as the medical concept gives us a higher *match* percentage than trying to match CUIs. Importantly, the dependence on unstructured data for resolution of temporal eligibility criteria is higher than structured data. There is especially a huge gap between the structured and unstructured data match in the case of prostate cancer, where structured data only contributes to the resolution of 9% of the criteria.

	CLL		Prostate Cancer	
	CUI	Medical Concept	CUI	Medical Concept
Structured Data Match	35%	37%	9%	9%
Unstructured Data Match	53%	59%	75%	77%

Table 2: Eligibility Criteria-level Analysis on CLL and Prostate Cancer Trials and Patient Records

	CLL	Prostate Cancer
Cross-Narrative Temporal Reasoning	33%	35%
Information fusion $L = S_k \cap U_j$ is not empty	24%	3%
Information fusion $L = S_k \cap U_j$ is empty	17%	1%

Table 3: Eligibility Criteria that require Cross-narrative Temporal Reasoning and Information Fusion for resolution

We observed that from the temporal criteria requiring unstructured data for resolution, frequently intra-narrative temporal reasoning was sufficient for resolving temporal constraints. The learned time-bins, along with the admission and discharge dates on each narrative, were useful in assigning medical concepts to coarse time-periods and in resolving 41% of the eligibility criteria that required an unstructured data match. For instance, the constraint, "patients with a *distant history* (*greater than 6 months before* study entry) of venous thromboembolic disease are eligible", requires mapping of venous thromboembolic disease to a time-bin *way before time*. Whereas "clinically significant bleeding event *within the last 3 months*, unrelated to trauma, or underlying condition that would be expected to result in a bleeding diathesis" required fine-grained temporal ordering of medical concepts.

Further, as shown in Table 3, from the criteria that required unstructured data for resolution, 33% and 35% required cross-narrative temporal reasoning in the CLL and prostate cancer dataset respectively. A criteria such as, "fever > 100.5°F for 2 weeks without evidence of infection", requires extracting the fact that fever lasted for 2 weeks by examining multiple mentions of fever across history and physical reports and discharge summaries to determine when fever started and stopped. This additionally requires the ability to perform coreference resolution across clinical narratives.[19] Criteria requiring information from both structured and unstructured data (information fusion) were determined based on the presence of the medical concepts in the criteria across these data sources. For instance, "if they have achieved stable blood pressure (bp) on a regimen of over 2 drugs after *6-8 weeks* of therapy." The value of bp can be obtained from the structured data, however the nuanced relationship information about the drug regimen that was prescribed to stabilize bp, along with its time duration, requires time-bin learning and cross-narrative temporal reasoning.
We observed that while a large percentage of CLL criteria required fusion, the lower number of prostate cancer criteria is mainly due to limited structured data available for prostate cancer.

Discussion

We studied two datasets of patients – CLL and prostate cancer – and evaluated the usefulness of structured vs. unstructured data in recruiting for corresponding clinical trials. We observed that the type of structured data, its granularity, and the information available vary across patient datasets. While the CLL patient dataset has detailed structured data in the form of diagnoses lists, encounters list, procedures and lab values, the prostate cancer dataset has limited structured data mostly consisting of medication lists and lab values. More fundamentally, the data heterogeneity reflects the underlying tumor heterogeneity at multiple levels. These levels include: (1) patient referral patterns (2) patterns of disease treatment (3) and differences in disease stages. At The OSU James Cancer Hospital, the majority of prostate cancer patients tend to be referrals from community oncologists or urologists after failure of first and second line therapies. In contrast, CLL patients are mostly evaluated from time of diagnosis and thus their entire case history is within the OSU system. Secondly, laboratory values for prostate cancer patients are often drawn at their local laboratory and subsequently faxed to their oncologist at OSU. These labs are not directly accessible and are found in the unstructured component of the medical record. In stark contrast, CLL labs are nearly universally drawn at OSU.

These tumor type differences would help explain our findings that prostate cancer requires the use of the unstructured data more frequently. The end result is that prior treatment history for prostate cancer patients who are seen at a later stage will have their disease course and treatment course summarized in the unstructured narrative. CLL patients are captured at an earlier stage and therefore their disease course and treatment history is more easily obtained from the structured text. This tumor type heterogeneity is reflected in the diagnosis codes that are available. In the case of CLL, these codes are useful in checking eligibility criteria that check for the presence or absence of a medical condition can be resolved easily from the structured data using these lists. In case of prostate cancer, this data is not as complete.

Tumor heterogeneity aside, structured data may also fail if the medical concept is at a finer level of granularity than what is required for an exact match. In such cases, examining the unstructured data for additional information, or additional processing to check for related higher level concepts for medical events in the structured data may help better resolve the eligibility criteria.

Conclusion

We performed an empirical evaluation of clinical trial eligibility criteria resolution using structured and unstructured patient datasets from CLL and prostate cancer. We observed that unstructured data is essential to resolving eligibility criteria in 59% of the CLL trial criteria and 77% of the prostate cancer trials. We also demonstrated the need for cross-document temporal relation learning and information fusion across structured and unstructured data sources. Although structured data is useful in resolving certain criteria, it is limited by information granularity and structured data type. Thus, structured data is best used for first pass filtering of EHR data in eliminating a criterion based on the presence or absence of a certain lab test or diagnoses, prior to a more nuanced second pass using unstructured data. Moreover, improving the coverage of the structured data in the EHR would improve its ability to be used as a clinical trial recruitment tool.

Acknowledgements

The project described was supported by Award Number Grant R01LM011116 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

References

1. Demner-Fushman, D., Chapman, W. W., & McDonald, C. J. (2009). What can natural language processing do for clinical decision support?. *Journal of biomedical informatics*, *42*(5), 760-772.
2. Tange, H. J., Schouten, H. C., Kester, A. D., & Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association*, *5*(6), 571-582.
3. Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association*, *18*(5), 540-543.
4. Li, L., Chase, H. S., Patel, C. O., Friedman, C., & Weng, C. (2008). Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. In *AMIA Annual Symposium Proceedings*(Vol. 2008, p. 404). American Medical Informatics Association.
5. Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., & Johnson, K. B. (2011). Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, *18*(2), 181-186.
6. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, *35*, 128-44.
7. Wang, X., Hripcsak, G., Markatou, M., & Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, *16*(3), 328-337.
8. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, *17*(5), 507-513.
9. Boland, M. R., Tu, S. W., Carini, S., Sim, I., & Weng, C. (2012). EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria.*AMIA Summits on Translational Science Proceedings*, *2012*, 71.
10. Ross, J., Tu, S., Carini, S., & Sim, I. (2010). Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits on Translational Science Proceedings*, *2010*, 46.
11. Zhou L, Hripcsak G. Temporal reasoning with medical data – A review with emphasis on medical natural language processing. J Biomed Inform. 2007; 40(2):183-202.
12. Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*.
13. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., and Setzer, A. (2008). Semantic Annotation of Clinical Text: The CLEF Corpus. In Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining, pages 19–26. 3, 32, 39.
14. Carlo, L., Chase, H. S., & Weng, C. (2010). Aligning structured and unstructured medical problems using UMLS. In *AMIA Annual Symposium Proceedings* (Vol. 2010, p. 91). American Medical Informatics Association.
15. Köpcke, F., Trinczek, B., Majeed, R. W., Schreiweis, B., Wenk, J., Leusch, T., Prokosch, H. U. (2013). Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC Med. Inf. & Decision Making*,*13*, 37.
16. Aronson, A. R. (2006). Metamap: Mapping text to the UMLS metathesaurus. *Bethesda, MD: NLM, NIH, DHHS*.
17. Raghavan, P., Fosler-Lussier, E., & Lai, A. M. (2012, June). Temporal classification of medical events. In Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (pp. 29-37). Association for Computational Linguistics.
18. Raghavan, P., Fosler-Lussier, E., & Lai, A. M. (2012, July). Learning to temporally order medical events in clinical text. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 70-74). Association for Computational Linguistics.
19. Raghavan, P., Fosler-Lussier, E., & Lai, A. M. (2012, June). Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 731-741). Association for Computational Linguistics.

Modeling Clinical Context: Rediscovering the Social History and Evaluating Language from the Clinic to the Wards

Colin Walsh, MD, Noémie Elhadad, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

Social, behavioral, and cultural factors are clearly linked to health and disease outcomes. The medical social history is a critical evaluation of these factors performed by healthcare providers with patients in both inpatient and outpatient care settings. Physicians learn the topics covered in the social history through education and practice, but the topics discussed and documented in real-world clinical narrative have not been described at scale. This study applies large-scale automated topic modeling techniques to discover common topics discussed in social histories, to compare those topics to the medical textbook representation of those histories, and to compare topics between clinical settings to illustrate differences of clinical context on narrative content. Language modeling techniques are used to consider the extent to which inpatient and outpatient social histories share in their language use. Our findings highlight the fact that clinical context and setting are distinguishing factors for social history documentation, as the language of the hospital wards is not the same as that of the ambulatory clinic. Moreover, providers receive little feedback on the quality of their documentation beyond that needed for billing processes. The findings in this study demonstrate a number of topics described in textbooks – schooling, religion, alternative health practices, stressors, for example - do not appear in social histories in either clinical setting.

Introduction

Increasing attention has been paid in the last decade to the importance of social and cultural factors with respect to health and disease. The Institute of Medicine in 2006 released its report, "Genes, Behavior, and the Social Environment", which called for "transdisciplinary, collaborative research" regarding "key social variables" affecting health outcomes such as education, income, social networks, and work conditions.[1] Behavioral, social and environmental factors have been linked to a panoply of issues including overall mortality, chronic diseases such as heart failure, and mental disorders such as rates of suicide.[2–5]

The routine history and physical examination in both inpatient and ambulatory settings is intended to query patients about social, behavioral, and environmental factors. Physicians in training learn the requisite components of a complete social history from medical textbooks on history-taking and the physical exam (Figure 1).[6] But the quality and comprehensiveness of these histories vary in clinical practice. Social histories can be collected in a heterogeneous manner depending on clinical setting and provider-patient rapport. Actual documentation of those histories can range from structured, coded data to electronic free-text to pen and paper charting.

Personal and Social History

- **Occupation**
- **Last year of schooling**
- **Home situation and significant others**
- **Sources of stress, both recent and long term**
- **Important life experiences, such as military service**
- **Leisure activities**
- **Religious affiliation and spiritual beliefs**
- **Activities of daily living (ADLs)**
- **Lifestyle habits**
  - **Exercise and diet**
  - **Safety measures**
  - **Alternative health care practices**

Figure 1. Textbook components of personal and social history.

Assessing quality of documentation remains a challenge to conduct at scale. While physicians are taught to conduct comprehensive histories to capture a holistic view of patient wellness, the gold standard is an individual-level chart

review with either the patient or physician as unit of analysis. Methods to permit large-scale semi-automated assessment of clinical narrative could have implications in the domains of quality of care and of practice improvement. Both clinical housestaff trainees and licensed practitioners who have moved on from the supervisory environment of residency programs might benefit from such tools. Documentation of social history, both because it relies on narrative and it has no established best-practices outside of the textbook guidelines, is in need of robust, scalable content-analysis methods.

The underlying research questions driving this study stem from the need to understand the common topics conveyed in a large collection of real-world social histories as well as the parity of those histories to "textbook teaching" in medical education. A related question considers what – if any – qualitative and quantitative differences exist between the language of social histories taken in inpatient and outpatient clinical settings

The internal medicine service at Columbia University Medical Center incorporates both inpatient hospital wards and ambulatory clinic settings and as such provides an opportunity to answer these research questions at scale. Both clinical contexts comprise hundreds of physicians with thousands of patient encounters per year. Techniques of topic modeling and of language modeling are used to quantify differences between those settings. Topic modeling is an attractive approach to discover aspects of social histories in a large collection of notes; it is unsupervised and as such remains robust to the idiosyncrasies of clinical language; it does not need any gold-standard annotation. Language modeling permits quantification of the extent to which inpatient and outpatient social histories differ in their lexical choice, again, in an unsupervised and robust fashion.

Background

Prior studies have considered the content of social histories in clinical text in public datasets and in small samples of clinical notes.[7,8] Melton et al performed content analysis of social and behavioral history datasets in the public health domain and showed an emphasis on smoking, alcohol use, drug use, and employment data.[7] Chen et al performed manual evaluation of the content of a small number of social histories contained in clinical notes and mapped them to HL7 and openEHR standards.[8] Our study is situated in the framework of unsupervised learning from a large corpus of clinical notes through both topic and language modeling.

Topic modeling is an established method, which, given a collection of documents and a pre-defined number of topics (K), identifies ranked lists of words (topics) according to which the documents can be described.[9,10] It operates as a type of dimensionality reduction from the entire vocabulary of documents to the K topics. Topic modeling has been widely used in non-clinical fields of natural language processing (NLP) and is gaining purchase in clinical natural language processing.[11–15] While these methods have been applied to clinical text, they have not been applied to the task of topic modeling of social histories in free-text clinical notes.

Language modeling is an older technique applied in multiple areas of language technology.[16] Given a training set, it assesses the likelihood of a new string of text (e.g., a sentence or a document) in a testing set. By comparing the likelihood of one text – one test set – against language models derived from different training sets, it is possible to determine the combination of training set and test set that are closest in lexical pattern.

Methods

Dataset

After obtaining approval of the Institutional Research Board, a dataset was collected of electronic clinical text taken from the electronic health record at Columbia University Medical Center from 2005-2009 for inpatient documentation and from 2008-2009 for outpatient documentation. The disparity between years relates to the increased use of the "Primary Provider Clinic Note" Type following 2008. No other clinic note type clearly identified the author as primary care provider, and this identification was necessary for the intended comparison between admitting inpatient physician and outpatient primary care provider. Two note types were extracted: "Medicine Admission Notes", the preferred electronic note completed by residents and attending physicians on the internal medicine service during the study period; "Primary Provider Clinic Note", the preferred note used by ambulatory care physicians who identified themselves as primary providers for that encounter. Both note types were written in a single blank text box without section headings or coded data entry. Any section headings in the notes were written by providers including abbreviations and shorthand (e.g. "Social History", "SocHx", "SH").

To avoid the bias introduced by including multiple admission notes from a single patient, only the first admission note for each unique medical record number was included in the test corpus. This bias has previously been described.[11,15] Similarly, the phenomenon of "copy-paste" or "copy-forward" in which providers copy text from previously written notes was avoided by the selection of the initial admission note for each unique patient record.

Data Preprocessing
The clinical text was imported into the R statistical environment and relevant admission and primary provider notes were identified. Free-text entries (entire admission and primary provider notes) were parsed into XML with section headings corresponding to content headings in notes, e.g. "History of Present Illness", "Medications", "Physical Exam". The social history sections were extracted from the XML output into a single file and then parsed further at the level of sentences.[17,18] All text was normalized to lowercase, and punctuation was removed. Overall, the dataset consisted of individual social history sections with sentence boundary information.

Topic Modeling
Topic modeling assumes that a document is a mixture of topics, and a topic is a mixture of words. The study goal was to capture the aspects of social histories that are common across a large population of patient records. To coax the topic modeling towards discovering these aspects, sentences in social history sections were considered to be individual documents. The hypothesis is that a sentence is a mixture of topics to be discovered.
Stop words and artifactual characters ("h" as part of "h o" or "history of") were removed in formatting text for latest Direchlet Allocation (LDA). Two sub-corpora were created – all inpatient social histories 2005-2009 and all outpatient social histories 2008-2009 – and LDA was applied to each independently.

A practicing internal medicine physician reviewed each set of topics manually. Topics were mapped whenever possible to one or several textbook topics (as described in Figure 1).

Perplexity Analysis of Inpatient- and Outpatient-Derived Language Models
Topic modeling allowed identification of aspects of inpatient and outpatient social histories. Language modeling and perplexity analysis enabled quantification of whether the language of the two settings differ (even if they cover similar topics). Perplexity invokes the concept of cross entropy to compare two different language models using a test set because the true underlying probability distribution in a corpus is unknown. Cross entropy is a step to quantifying the uncertainty in modeling language in a corpus imperfectly; the lower the cross entropy (and the perplexity which relies on it), the better the fit of the model to the data.

Two years of clinical text data were isolated for perplexity analysis; one training set of inpatient text was constructed from 2008 inpatient social histories, and one training set of outpatient text was constructed from 2008 outpatient social histories. The inpatient and outpatient social histories from 2009 comprised the two testing sets. Two N-gram language models were trained, one inpatient and one outpatient. Perplexity was calculated within and between clinical settings, e.g. the inpatient language model was tested against both inpatient and outpatient testing sets. The lower the perplexity, the better the lexical fit between unseen text and the training set. Thus, when comparing the perplexity of one corpus against two different models, the set with the lowest perplexity represented the training set that matched the unseen text more closely.

Experimental Setup
For experiments with topic modeling, variational inference implementation of LDA was used.[19] The default parameters were used including random topic initialization, variational inference set to converge algorithmically, and a maximum of one hundred iterations of expectation maximization to estimate hyperparameters of $\propto$ and β. The number of topics was set manually and various numbers of topics were tested: ten, twenty, thirty. Results are reported for ten topics only.

For experiments with language modeling, the Stanford Research Institute Language Modeling (SRILM) toolkit was used with a trigram model with Kneser-Ney smoothing.[20,21]

Results

From 2005-2009, the clinical data repository contained 64,610 notes of type "Medicine Admission Note". After removal of duplicate medical record numbers (implying subsequent admissions) and text parsing to identify sections, the Inpatient Social History Corpus was composed of 48,944 documents. From 2008-2009, the repository contained 15,154 notes of type "Primary Provider Clinic Note". After removal of duplicates and parsing, the Outpatient Social History Corpus contained 7,796 documents.

Topic Modeling

Table 1. Inpatient Social History Topic Models

Topic	Word Clusters (rank order of words, ten words shown)	Topic Label (manually assigned)
1	etoh, drugs, denies, not, yrs, but, home, tob, never, worked	Lifestyle Habits, Employment
2	live, ny, ppd, nl, boyfriend, stds, remote, still, new, Washington	Home Situation, Sexual History
3	works, wife, last, us, marijuana, heavy, only, grandchildren, weekends, prev	Family, Lifestyle Habits
4	lives, quit, her, husband, illicits, since, son, pt, alcohol, mother	Family, Lifestyle Habits
5	use, now, work, independent, former, factory, none, age, history, old	Employment
6	dr, ago, sexually, working, adls, one, here, illicit, occasional, drinks	Background, Employment, Lifestyle Habits
7	tobacco, years, daughter, active, currently, alone, drug, smoked, past, she	Smoking (Lifestyle Habit)
8	no, children, came, worked, but, beer, this, iadls, clear, vices	Family, Lifestyle Habits, Activities of Daily Living
9	day, gt, retired, moved, disability, does, ivdu, sexual, school, months, two, uses	Employment, Support, Lifestyle Habits, Education
10	smoking, used, social, nyc, who, cocaine, stopped, family, separated, alone	Lifestyle Habits (smoking, drugs), Social Support

Table 2. Outpatient Social History Topic Models

Topic	Word Clusters (rank order of words, ten words shown)	Topic Label (manually assigned)
1	now, work, home, pt, adls, independent, kids, all, care, iadls	Family, Activities of Daily Living
2	not, active, currently, sexually, working, last, hx, does, hiv, sexual	Sexual History
3	denies, use, drug, alcohol, illicit, other, habits, up, wt, bp	Lifestyle Habits
4	children, dr, her, married, here, born, moved, nyc, sister, living	Family, Support
5	years, ago, quit, yrs, smoking, social, smoked, occasional, ppd, cocaine	Lifestyle Habits (smoking, illicit drugs), Preventive Testing
6	wife, worked, since, us, hha, retired, previously, his, factory, came	Employment
7	but, she, one, he, who, time, separated, old, well, father	Family
8	lives, daughter, works, husband, alone, son, mother, unemployed, two, Bronx	Family, Employment

9	day, used, never, past, gt, former, year, week, per, disability	Modifiers
10	no, etoh, drugs, tobacco, illicits, tob, ted, cigs, rare, rrr	Lifestyle Habits

Results of topic modeling are presented Tables 1 and 2. The inpatient social history topics included multiple representations of lifestyle habits – specifically smoking, alcohol use, illicit drug use. The outpatient social history topics also included lifestyle habits as well as topics more clearly composed of family, social support, sexual history, and employment.

In comparison to the textbook version of the social history presented in Table 1, it is clear that lifestyle habits – tobacco, alcohol, illicit drug use, specifically – are the most commonly reflected aspects of social history within inpatient clinical text. Inpatient social histories allude to topics of family, social support, and education, for example, but words in these categories are mixed with other aspects of the social history. The outpatient social histories demonstrate aspects in common with inpatient histories, but the aspects are more clearly defined and more likely to be topics composed entirely of a single aspect. Outpatient Topic #2 for example includes multiple terms related to sexual history and HIV testing while Inpatient Topic #2 mingles words associated with sexual history ("boyfriend", "stds" for sexually transmitted diseases) with other aspects including preventive testing ("ppd") and living environment ("lives", "ny" for New York). Textbook topics that were not apparent in either inpatient or outpatient social histories include: schooling, religious affliation, significant stressors, major life experiences, alternative health practices.

Perplexity Analysis of Inpatient and Outpatient Social History Language Models

Table 3 summarizes the results of perplexity calculation using training sets of inpatient 2008 and outpatient 2008 social history text on testing sets of inpatient 2009 and outpatient 2009 social history text. Out-of-vocabulary words (OOVs) are included as are the numbers of sentences and words in each training corpus.

OOVs from inpatient to outpatient settings are much larger across settings than within the same setting (60K unknown words vs. 26K) and, similarly, inpatient social histories are better modeled by the inpatient language model than the outpatient one (202.9 vs. 331.9 perplexity). Language of outpatient social histories is also better modeled according to an outpatient language model (123.3 vs. 385.8 perplexity). This fact remains even though the number of OOVs in the outpatient test set outnumbers the OOVs in the inpatient set when tested by the outpatient language model (18K vs. 16K).

Table 3. Testing across clinical contexts (inpatient to outpatient and vice versa) reveals higher perplexity.

Training Set	Testing Set	# Sentences in Training Set	# Words in Training Set	OOVs in Testing Set	Perplexity of Testing Set
Inpatient 2008	Inpatient 2009	51,978	373,390	26,075	202.9
Inpatient 2008	Outpatient 2009	51,978	373,390	60,767	385.8
Outpatient 2008	Inpatient 2009	29,764	219,513	16,805	331.9
Outpatient 2008	Outpatient 2009	29,764	219,513	18,175	123.3

Discussion

The principal findings of this study demonstrate that large-scale topic modeling can reveal topics of interest within a chosen section of clinical text. Inpatient social histories correspond to topics of interest to physicians in an intermediate acuity clinical setting. Lifestyle habits – particularly tobacco, alcohol, and illicit drug use – as well as basics of employment and family support are frequently queried areas within social histories. However, these topics in inpatient text are frequently mixed within sentences and are therefore not clearly demarcated. Outpatient social history text, on the other hand, reveals more clearly delineated topic areas as well as more complete social histories being taken by primary care providers. This result corresponds intuitively with clinical teaching and with an

encounter in a lower acuity care setting in which continuity of care is more obviously anticipated. Both topic models failed to reveal some topics emphasized as important to a holistic view of a patient's wellness. The social history topics did not include education or schooling necessary to approximate health literacy, religious affiliation that might inform patients' treatment decisions or end-of-life choices, or alternative health practices that may affect patients' health overall (e.g. alternative weight loss therapies or alternative therapeutics that might interact with prescribed medications). It remains an open question as to whether clinical histories should reflect the textbook closely or whether clinical workflows in practice are appropriately different than traditional teaching.

The perplexity analysis reveals another important finding. The language of the wards is not the same as that of the clinic. Language models derived from a particular clinical context are prone to more entropy and higher computation requirements when they are applied across care settings. This result has important implications for language modeling as clinical context should be considered in the development and evaluation of novel tools in this space. Applications of these findings could fall into the domains of documentation quality, medical education, or billing. Utilizing content modeling through LDA and language models such as those described here might enable physicians to identify and correct deficiencies in clinical notes for quality purposes or to maximize the level of billing for the exams that they perform. Medical students could be prompted to ask patients about as-yet undocumented topics in notes as they learn to take comprehensive histories. Clinical context should be taken into account to optimize the performance of requisite language models in all of these areas.

Strengths of this study include a large amount of real-world clinical text as well as the use of entirely free-text documentation as the original dataset. It demonstrates the strength of automated methods to assign and extract sections from otherwise unformatted blocks of text in which a variety of abbreviations and shorthand are common (e.g. "Social hx", "SocHx", "SH"). The use of unsupervised methods, which do not assume any semantic information about the words in the text, enabled identification of highly meaningful words with respect to social history that would otherwise be difficult to extract from text automatically. For instance, "dr" in this analysis referred not to doctor, but to Dominican Republic, a common country of origin for many of the patients in this study population. While there was no gold standard available in this corpus, textbook information was leveraged to validate and compare the results of topic modeling.

Limitations of this study include a larger dataset for inpatient data than outpatient data as a result of the note types selected for study. Similarly, notes were selected from 2005-2009 and not more recent years because of changes in documentation practice in both care settings since 2009. Free-text documentation without templates or other coded data is now much less common, but such semi-structured data demonstrates less clearly the flexibility and robustness of the natural language processing methods used here. The noise of clinical data is another limitation common to all methods incorporating clinical text. The perplexity analysis was limited by the asymmetry of sizes of corpora for training – a result of the difference in number of visits between inpatient and outpatient settings. It is possible that some elements of social history were discussed outside of the sections labeled "Social History" by providers (such as History of Present Illness), but these were filtered out at the corpus processing stage. And because this analysis took place at corpus-level, there may be low-frequency relevant topics in individual notes outside of social histories sections; these were not the subject of this investigation.

This study extends the work of others by demonstrating that topics of clinical text can be rediscovered through unsupervised machine learning. It also demonstrates the computational impact of modeling language without concern for clinical context. The lower perplexity associated with testing language models within opposed to across clinical settings has important implications for subsequent endeavors in this discipline. Finally, it suggests a system of practice and documentation improvement could discover areas of deficiency in documentation in real-world clinical text. Providers do not currently have dedicated mechanisms of documentation quality review outside of that necessary to accomplish billing tasks. This method could be one step to the creation of such a system to help providers improve at a task they perform multiple times per day without feedback – documentation of the clinical encounter.

Future research can extend this technique to discover topics in other sections of the clinical note or in other note types entirely. Other provider types (surgeons, nurse practitioners, emergency medicine physicians) also incorporate different lexicons in their clinical narratives; these methods could be applied to those contexts to reveal aspects of that language. Language models can be employed for specific levels of care and care settings. Clinical experience

dictates that the language of the intensive care unit varies from the medical wards, for example, and this technique could be one step to verify and subsequently overcome that obstacle. The results of this study could also inform further data modeling including predictive analytics and the extraction of free-text data for incorporation into coded datasets.

Conclusion

Factors of social support, cultural facets, and behaviors are important components of the medical history collected in both inpatient and outpatient encounters. The high quantity of clinical text – often unstructured free-text – requires automated methods that can elucidate underlying aspects of those histories and pave the way for secondary use, quality assessment, clinical training, and decision support incorporating such data in a structured way. This study demonstrated the aspects of social history from both inpatient and outpatient clinical encounters. Inpatient encounters are typified by shorter social histories with an emphasis on immediately cogent behaviors such as smoking, alcohol use, and illicit drugs. Outpatient social histories include greater breadth and depth of topics covered, but neither type of social history touched on significant aspects of patient wellness such as schooling, religious affiliation, or life stressors. Perplexity analysis suggests strongly that clinical context and setting must be incorporated in the design of computational methods to process and benefit from the clinical narrative.

Acknowledgements

This work is supported by T15 LM007079 (CW) and National Library of Medicine award R01 LM010027 (NE).

References

1. Institute of Medicine. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. 2006.

2. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. JAMA : the journal of the American Medical Association. 2004 Mar 10;291(10):1238–45.

3. Kleiman EM, Liu RT. Social support as a protective factor in suicide: Findings from two nationally representative samples. Journal of affective disorders. Elsevier; 2013 Mar 2

4. De la Cámara AG, Guerravales JM, Tapia PM, Esteban EA, Del Pozo SVF, Sandubete EC, et al. Role of biological and non biological factors in congestive heart failure mortality: PREDICE-SCORE: a clinical prediction rule. Cardiology journal. 2012 Jan;19(6):578–85.

5. Rasic DT, Belik S-L, Elias B, Katz LY, Enns M, Sareen J. Spirituality, religion and suicidal behavior in a nationally representative sample. Journal of affective disorders. 2009.

6. Bickley LS, Szilagyi PG, Bates B. Bates' guide to physical examination and history-taking / Lynn S. Bickley, Peter G. Szilagyi. 11th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013. p. xxv, p. 994

7. Melton GB, Manaktala S, Sarkar IN, Chen ES, Health C. Social and Behavioral History Information in Public Health Datasets. Proceedings of the AMIA Annual Symposium. 2012. pp. 625–34.

8. Chen ES, Manaktala S, Sarkar IN, Melton GB. A Multi-Site Content Analysis of Social History Information in Clinical Notes. Proceedings of the AMIA Annual Symposium. 2011. pp. 227–36.

9. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993–1022.

10. Blei D. Probabilistic Topic Models. Communications of the ACM. 2012 Nov;77–84.

11. Arnold C, Speier W. A Topic Model of Clinical Reports. 2012;90024:1031–2.

12. Salleb-aouissi A, Radeva A, Passonneau RJ, Xie B, Khattak FK, Tomar A, et al. Diving into a Large Corpus of Pediatric Notes. Proceedings of the ICML Workshop on . 2011;

13. Perotte A, Bartlett N, Elhadad N, Wood F. Hierarchically Supervised Latent Dirichlet Allocation. Proceedings of the Neural Information Processing Systems Conference (NIPS). 2011;1–9.

14. Halpern Y, Horng S, Nathanson LA, Shapiro NI. A Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text. Proceedings of the ICML Workshop on Clinical Data Analysis. 2012;

15. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC bioinformatics. 2013 Jan 16. 14(1):10.

16. Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. May 1999.

17. Li Y, Lipsky Gorman S, Elhadad N. Section classification in clinical notes using supervised hidden markov model. Proceedings of the ACM international conference on Health informatics (IHI). 2010. pp. 744.

18. Tsuruoka Y, Tateishi Y, Kim J, Ohta T. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Lecture Notes in Computer Science. 2005. pp. 382–92.

19. Blei DM. Latent Dirichlet Allocation in C. Available from: http://www.cs.princeton.edu/~blei/lda-c/

20. Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Computer Speech & Language. 1999 Oct;13(4):359–93.

21. Wang W. SRILM at Sixteen : Update and Outlook. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. 2011.

Longitudinal Analysis of New Information Types in Clinical Notes

Rui Zhang, PhD[1], Serguei Pakhomov, PhD[1, 2], Genevieve B. Melton, MD, MA[1,3]
[1]Institute for Health Informatics; [2]College of Pharmacy; [3]Departments of Surgery
University of Minnesota, Minneapolis, MN

Abstract
It is increasingly recognized that redundant information in clinical notes within electronic health record (EHR) systems is ubiquitous, significant, and may negatively impact the secondary use of these notes for research and patient care. We investigated several automated methods to identify redundant versus relevant new information in clinical reports. These methods may provide a valuable approach to extract clinically pertinent information and further improve the accuracy of clinical information extraction systems. In this study, we used UMLS semantic types to extract several types of new information, including problems, medications, and laboratory information. Automatically identified new information highly correlated with manual reference standard annotations. Methods to identify different types of new information can potentially help to build up more robust information extraction systems for clinical researchers as well as aid clinicians and researchers in navigating clinical notes more effectively and quickly identify information pertaining to changes in health states.

Introduction
Electronic health record (EHR) systems provide significant opportunities to integrate and share health information and increase the efficiency of health care delivery, as well as re-use clinical data for research studies. Of the many functionalities of EHRs, clinical note documentation is an essential part of patient care. A large number of efforts in clinical research have focused on identification of patient cohorts that meet clinical eligibility criteria for studies, with some methods aiming to extract these criteria from both structured data and unstructured texts including phenotype extraction[1, 2] and drug related information extraction[3, 4]. Structured information from billing and administrative codes (i.e., ICD and CPT codes), however, is often insufficient to accurately find patient cohorts for clinical research[5]. Unstructured text of clinical notes often contains the necessary more detailed information but information extraction (IE) systems based on clinical texts alone often have difficulty achieving adequate performance[6, 7]. For instance, when ICD-9 codes and IE from clinical texts are combined, a higher accuracy for detecting colorectal cancer can be achieved[8].

Most EHR clinical documentation modules allow text from one note to be reused in subsequent notes ("copy-and-paste"). The practice of copying information from previous documents and pasting into the current clinical note being constructed is often used to shorten the time spent documenting. However, one of the unintended consequences of frequent copying and pasting of patient data especially with complicated care episodes or clinical courses is that copy-and-paste can create large amounts of replicated information resulting in longer and less readable notes than those seen previously with paper charts[9-11].

Clinical texts with significant amounts of redundant information combined with large numbers of notes not only increase the cognitive burden and decision-making difficulties of clinicians[11-16], but also may impact the accuracy and efficiency of IE systems[17]. Moreover, redundant information can also contain a mixture of outdated information or errors in the information copied making it difficult for clinicians to interpret information in these notes most effectively[12]. It has been suggested that considering the structure of texts and redundancy before implementing IE tasks may be valuable[17, 18]. Thus, effective classification of redundant and new information could potentially help to improve the performance of the IE systems for clinical research.

Large amounts of redundant information have been found in both inpatient and outpatient notes with automated methods[12, 19, 20]. Hammond et al. performed pair-wise comparisons to detect identical word sequence and found 12.5% of information in notes copied[12]. Wrenn et al. used global alignment techniques to quantify redundancy in inpatient clinical notes,[19] finding large amounts of redundant information which increased over the course of a patient's inpatient stay. Zhang et al. modified the Needleman-Wunsch algorithm, a classic global sequence alignment technique used in bioinformatics, to quantify redundancy in outpatient clinical notes with similar findings[20].

Research has also focused upon methods to identify relevant new information and evaluation of the potential impact of redundant information on clinical practice. One of the recognized gaps in these approaches is that these methods do not intrinsically provide more details about the types of new information (e.g., medication, disorders, symptoms).

Categorization of new information may aid clinicians and researchers in finding specific types of new information more easily in a more purposeful manner within notes. The objective of this study was to extract specific types of relevant new information, specifically problem/disease (or comorbidities), medication, and laboratory.

Methods

The three-part methodological approach for this study included: 1) developing a reference standard of new information with information type; 2) identification of new information using an *n*-gram modeling technique modified for clinical texts; and 3) extraction of semantic types and key terms from identified new information.

Data collection

Outpatient EHR notes were retrieved from the University of Minnesota Medical Center affiliated Fairview Health Services. For this study, we randomly selected 100 geriatric patients with multiple co-morbidities, allowing for relatively large numbers of longitudinal notes in the outpatient clinic setting. To simplify the study, we limited the notes to office visit notes arranged chronologically. These notes were extracted in text format from the Epic™ EHR system between 06/2005 and 06/2011. Institutional review board approval was obtained and informed consent waived for this minimal risk study.

Automated methods to identify new information semantic types

We used a hybrid method with *n*-gram models and heuristic information previously developed to identify new information in clinical documents[21]. In brief, after text pre-processing, *n*-gram models with classic and TF-IDF stopword removal, lexical normalization, and heuristic rules were used to remove note formatting and adjustments by section. After obtaining new information within each note, this text was mapped to the UMLS[22] using MetaMap[23] with options to allow acronym/abbreviation variants (-a) and NegEx results (--negex). From this, we extracted semantic types using scores of 600 and over as the cutoff. To simplify the analysis, we restricted our detailed analysis to the specific types to identify information about problem/disease, medication, and lab results (Table 1)[22].

Table 1. Sections and semantic types for identifying category of new information.

Category	Semantic Types
Problem/Disease	[Disease or Syndrome], [Finding], [Sign or Symptom]
Medication	[Clinical Drug], [Organic Chemical, Pharmacologic Substance], [Biomedical or Dental Material]
Laboratory	[Laboratory Procedure], [Therapeutic or Preventive Procedure], [Diagnostic Procedure], [Amino Acid, Peptide, or Protein], [Biologically Active Substance]

Calculation of various types of new information proportion of patient notes

To calculate the NIP of each note, the new information algorithm was trained on previous *n* (e.g., 1, 2, …) notes to predict the new information of $(n+1)^{th}$ note for the whole corpus (100 patients). NIP was defined as the number of sentence (at least contain one piece of new information) divided by the total number of sentences of each note. We used the same method to further quantify NIP on the number (at a sentence or statement level) of various types of NIP for each note including disease (NDIP), medication (NMIP) and lab results (NLIP) based upon the identified semantic types of new information (Table 1). We then plotted NIP of various types for each patient over time. For the purposes of graphical display of notes temporally, we adjusted the dates of patient notes by a random offset of +/- 1 to 364 days.

Manually reviewed annotation as gold standard

Two medical intern physicians (aged 26 and 30) were asked to identify new and clinically relevant information based on all preceding documents chronologically for each patient using their clinical judgment. They were also asked to classify new information into the following types: problem, medication, laboratory, procedure or imaging, surgery history, family history, social history, and medical history. Longitudinal outpatient clinical notes from 15 patients with the last 3 notes for each patient annotated were selected for this study.

To maximize agreement, we first allowed the annotators to compare each other's annotations on a small separate corpus to reach a consensus on the standards for judging relevant new information. To assess inter-rater reliability, each physician later manually annotated a separate set of 10 overlapping notes. Cohen's Kappa statistic and percentage agreement were used to analyze agreement at a sentence or statement level.

In addition to the notes used to establish and measure inter-annotator agreement, 90 notes were annotated by medical interns. Forty notes were used for training and the remaining fifty notes for evaluation. Performance of

automated methods was compared to the reference standard and then measured for precision and recall at a sentence or statement level.

A third resident (3rd year) manually reviewed chronologically ordered 100 office visit notes from five individual patients to note new information in the format of short key terms. We then compared this with the automatically computed new information proportion (NIP) measure for each note, and extracted biomedical terms of various categories. Precision and recall for three types of new information within these notes were calculated.

Results

Annotation evaluation and method performance

Two raters showed a good agreement (which was improved from previous evaluations[21]) with identifying new information with Cohen's Kappa coefficient of 0.80 and percentage agreement of 97%. The precision and recall of the best method with *n*-grams are 0.81 and 0.84, respectively. The precisions values for extracting disease, medication, and laboratory new information are 0.67, 0.66, and 0.67; and the recall values are 0.72, 0.92, and 0.80, respectively.

Identification of various types of relevant new information

After calculating new information using the reference standards at the sentence level, we obtained the following percentage of various categories (e.g., lab, problem, medication) of relevant new information annotated by medical experts. The top three categories were problem (34.1%), medication (31.7%) and laboratory results (17.3%). Other types include procedures of imaging (5.0%), family history (2.8%) social history (2.7%), medical history (2.4%), surgery history (0.4%), and others (3.6%).

NIP, NDIP, NMIP, and NLIP were then calculated for each note where NIP = NDIP + NMIP + NLIP + NOIP. Note that NOIP represents other types of new information proportion (e.g., Mental Process). Individual patients were then selected and NIP, NDIP, NMIP, and NLIP plotted as illustrated in one patient in Figure 1 and Figure 2. Subjective new information for each note was also obtained for each note (Figure 1). Overall, notes with higher NIP correlated with more new information, and notes with lower NIP scores tended not to contain significant new information. Key biomedical concepts were extracted for each information category and were marked (using the automated new text extracted) for each note in Figure 2.

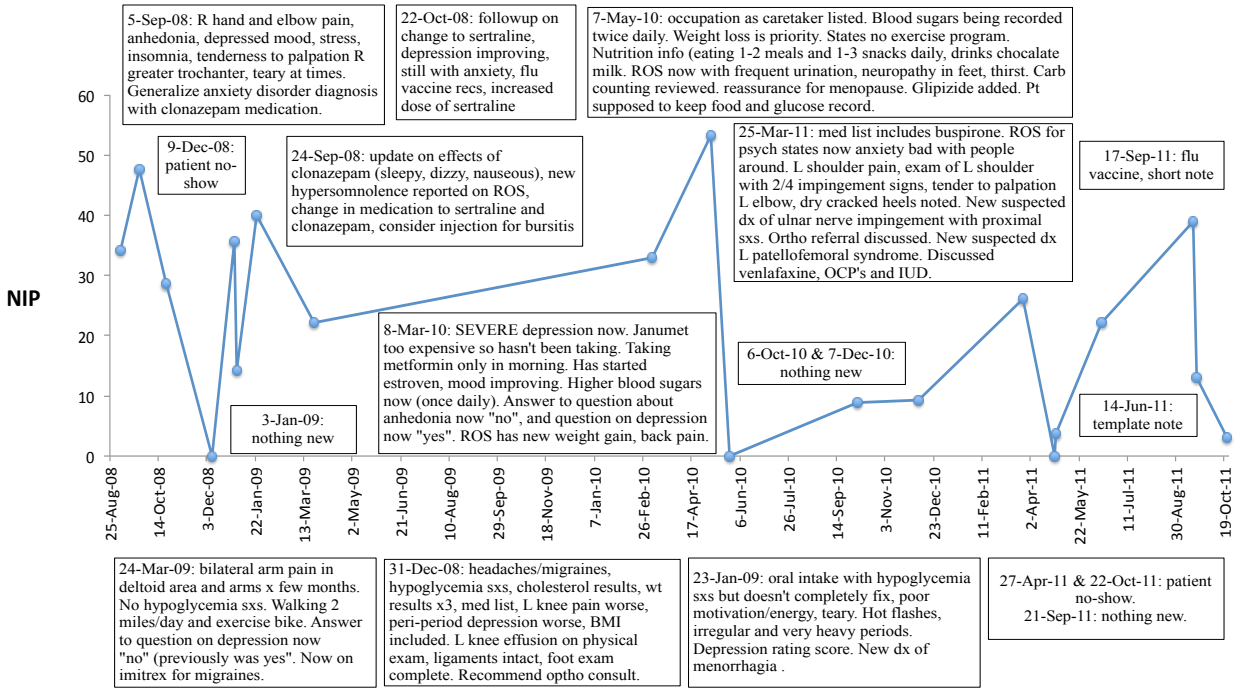


Figure 1. New information proportion (NIP) of clinical notes an illustrative patient. Boxes contain summarized new information.

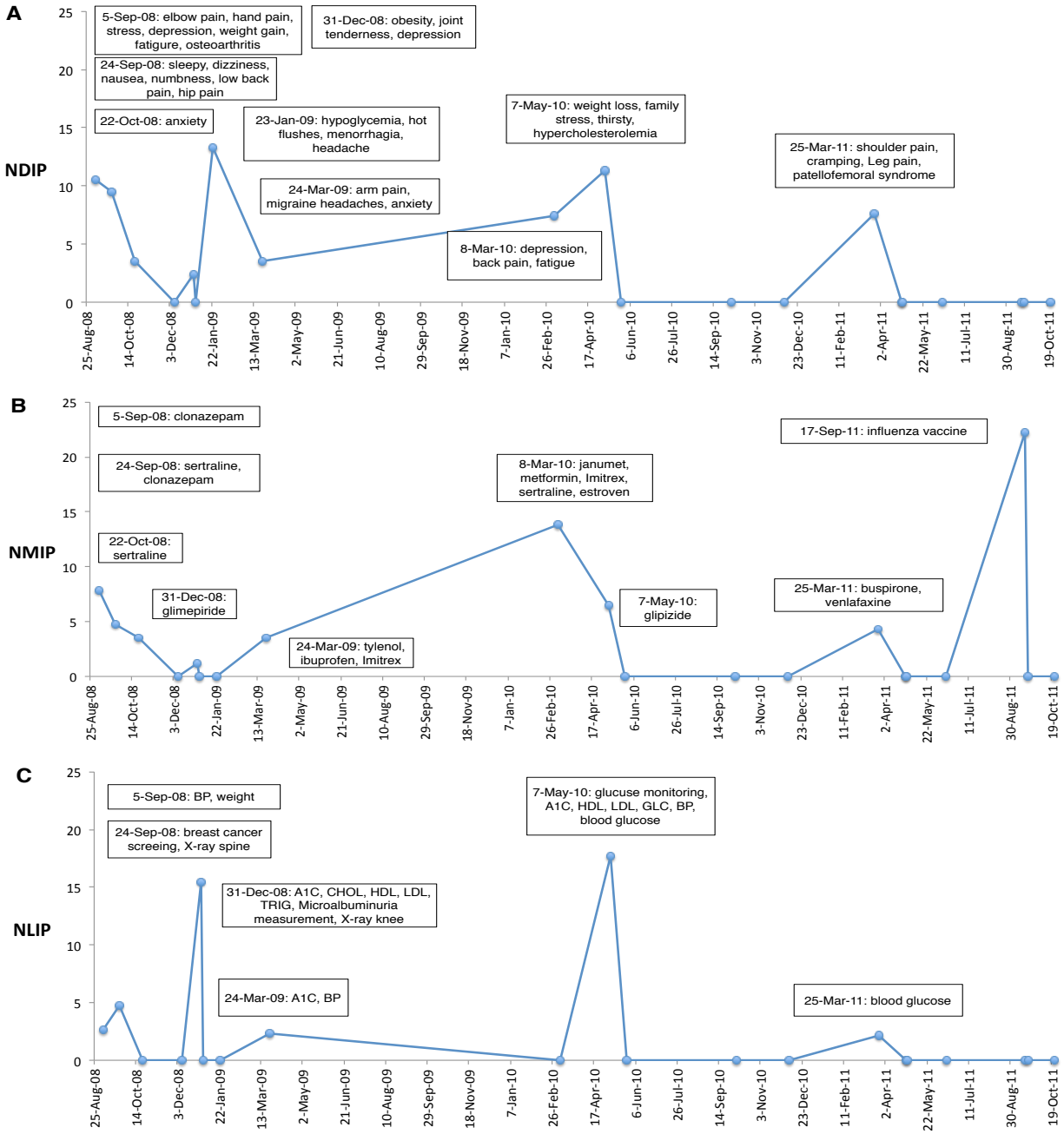


Figure 2. Plot of (A) NDIP (disease), (B) NMIP (medication), and (C) NLIP (laboratory) over time for the same patient as Figure 1. Biomedical concepts for each note included in boxes. NDIP, new problem/disease information proportion; NMIP, new medication information proportion; NLIP, new laboratory information proportion.

Discussion

In the field of clinical research informatics, many researchers focused on improvement of clinical NLP and data mining methods applied to EHR clinical texts, but minimal research has focused on the impact of the redundant nature of clinical texts it being ubiquitous and an issue for clinicians in reviewing a patient's notes. Cohen et al. recently reported that redundancy of clinical texts had impact on two text mining techniques: collocation extraction and topic modeling. The authors suggest examining the redundancy of a given corpus before implementing text mining techniques[17]. Thus, studies on automated methods to identify relevant new information represent a potential

set of techniques to improve the information extraction process from clinical notes. Previous work has demonstrated that NIP measures may be useful in identifying notes with clinically relevant new information[24]. While notes with higher NIP scores usually correlate with new findings, some other pertinent questions include answering questions such as "*Why are notes with high new information scores important?"* and "*What specific new information does this note contain?"*. This study examines types of new information in several important categories by dividing original NIP scores into various types of new information. Such classification of new information can potentially help clinical researchers navigate to specific types of information in clinical more effectively.

In comparing annotations based on UMLS concepts by medical residents to automated methods there were several key findings. We found some types of problem/disease information where automated methods identified information that was not included in the physician-generated reference standard. In one example, symptoms of *elbow pain*, *hand pain*, and *depression* were identified in the reference standard but several other symptoms such as *anhedonia* and *insomnia* were not identified but were new. In contrast, with medications, automated methods incorrectly identified some medications. For example, (Figure 4b) new medications of *clonazepam* (5-Sep-08), *sertraline* (24-Sep-08), *metformin* (8-Mar-10), *estroven* (8-Mar-10), *glipizide* (7-May-10), and *buspirone* (25-Mar-11) were found via automated methods and by our expert annotators, but the method incorrectly found "*janumet*" from the sentence "…janumet was too expensive, so she did not take it." (8-Mar-10). Although we used NegEx functionality in MetaMap to account for negation, our automated method did not effectively deal with the co-reference issue (it refers to janumet). Another example is "*vernlafaxine*" from the note (25-Mar-10) "…another future option may be to try venlafaxine". Here, the physician only recommended the medicine instead of prescribing it accounting for another false negative example. Finally, with respect to laboratory information, there were examples where the physician annotator did not mark laboratory data. One reason for this is that *glucose* and *hemoglobin A1C* tests are routine monitoring tests, and clinicians will not focus on that unless there are significant changes of the results. We also faced mapping issues with respect to acronyms for laboratory procedures. For example, we had to translate "A1C" to its full name "Hemoglobin A1C" to be recognized by MetaMap. In follow-up studies, we may provide more detailed information (e.g., if the value excess the normal range) other than just listing laboratory name to aid clinicians to pay more attention to the specific lab results with unexpected values.

Our method has certain other limitations. Mapping techniques such as that provided with MetaMap do not give additional types of information such as changes in dosage for specific drugs. We also did not solve other semantic level issues as mentioned previously in the discussion, such as co-reference. Also, although we compared our results with annotated reference sample patient notes, this reference standard was not built at the concept level per sentence whereas our method used concepts at the biomedical term level, which was readily available. Currently, we have only looked at three types of new information; other types of information such as *Mental Process* could be another valuable set of semantic types to explore. In future research, we will also consider the use of specialized modules such as MedEx[25] to extract more details associated with changes in medication use, other than just providing drug name.

Conclusion

We used the combination of language models and semantic types not only to identify new information, but also to extract key new information at the biomedical term level. We found that our ability to extract new key terms with our methods had good correlation with expert judgment. As these methods for new information detection are further developed, they can potentially help researchers avoid the biased clinical texts due to the redundant information and find the information type of interest. Moreover, it can also aid clinicians in finding notes and information within notes with more detailed types of new information, such as new diseases or medications.

Acknowledgments

This research was supported by the American Surgical Association Foundation Fellowship (GM), University of Minnesota (UMN) Institute for Health Informatics Seed Grant (GM & SP), Agency for Healthcare Research & Quality (#R01HS022085-01) (GM), National Library of Medicine (#R01LM009623-01) (SP) and the UMN Graduate School Doctoral Dissertation Fellowship (RZ). The authors thank Fairview Health Services for support of this research. The authors thank Dr. Janet T. Lee for annotation of the patient notes for this study.

References

1. Kullo IJ, Fan J, Pathak J, Sayoya GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc. 2010 Sep;17(5):568-74.

2. Kho AN, Pacheco JA, Peissig PL, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Sci Transl Med. 2011 Apr 20;3(79):79re1.
3. Tatonetti NP, Denny JC, Murphy SN, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther. 2011 Jul;90(1):133-42.
4. Wang XY, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. J Am Med Inform Assoc. 2009 May-Jun;16(3):328-37.
5. Birman-Deych E, Waterman AD, Yan Y, Nilasena DS, Radford MJ, Gage BF. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. Med Care. 2005 May;43(5):480-5.
6. Penz JF, Wilcox AB, Hurdle JF. Automated identification of adverse events related to central venous catheters. J Biomed Inform. 2007 Apr;40(2):174-82.
7. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc. 2008:404-8.
8. Xu H, Fu Z, Shah A, et al. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. AMIA Annu Symp Proc. 2011;2011:1564-72.
9. Markel A. Copy and paste of electronic health records: a modern medical illness. Am J Med. 2010 May;123(5):e9.
10. Hirschtick RE. A piece of my mind. Copy-and-paste. JAMA. 2006 May 24;295(20):2335-6.
11. Yackel TR, Embi PJ. Copy-and-paste-and-paste. JAMA. 2006 Nov 15;296(19):2315; author reply -6.
12. Hammond KW, Helbig ST, Benson CC, Brathwaite-Sketoe BM. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. AMIA Annu Symp Proc. 2003:269-73.
13. 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):61-7.
14. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. J Am Med Inform Assn. 2011 Mar;18(2):112-7.
15. Patel VL, Kaufman DR, Arocha JF. Emerging paradigms of cognition in medical decision-making. J Biomed Inform. 2002 Feb;35(1):52-75.
16. Reichert D, Kaufman D, Bloxham B, Chase H, Elhadad N. Cognitive analysis of the summarization of longitudinal patient records. AMIA Annu Symp Proc. 2010:667-71.
17. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics. 2013;14:10.
18. Downey D, Etzioni O, Soderland S. Analysis of a probabilistic model of redundancy in unsupervised information extraction. Artif Intell. 2010 Jul;174(11):726-48.
19. Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. J Am Med Inform Assoc. 2010 Jan-Feb;17(1):49-53.
20. Zhang R, Pakhomov S, MaInnes BT, Melton GB. Evaluating Measures of Redundancy in Clinical Texts. AMIA Annu Symp Proc. 2011:1612-20.
21. Zhang R, Pakhomov S, Melton GB. Automated Identification of Relevant New Information in Clinical Narrative. IHI'12 ACM Interna Health Inform Sym Proc. 2012:837-41.
22. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70.
23. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36.
24. Zhang R, Pakhomov S, Lee JT, Melton GB. Navigating longitudinal clinical notes with an automated method for detecting new information. Stud Health Technol Inform. 2013;192:754-8.
25. 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. 2010 Jan-Feb;17(1):19-24.