On behalf of AMIA and the TBI and CRI Scientific Program Committees, we are pleased to welcome you to the eleventh AMIA Informatics Summit!

The Informatics Summit serve as the primary forum to connect with leaders in the field of informatics who are advancing translational science at the nexus of bioinformatics and clinical research.

The meeting brings together translational scientists, data scientists, informatics researchers and practitioners from academia, industry, government and non-profit sectors to share knowledge and best practices, and to forge collaborations across boundaries.

Equally emphasizing discovery and impact, the Informatics Summit welcomed submission on the innovation, evaluation, and implementation of transformative concepts, methods, and technologies to accelerate translational science and precision medicine. It includes contributions to foundational concepts and methodologies as well as applications with public health impact. The TBI Summit focuses on innovative methods and novel discoveries that advance understanding of human disease and enable personalized and precision medicine, while the CRI Summit highlights translational research from bench to bedside, practice, and communities.

The following types of submissions were submitted:

- Papers
- Student Papers
- Podium Abstract Presentations
- Posters
- Panels
- Partnerships in Innovation

Moreover, the following awards were presented:

- Marco Ramoni Distinguished Paper Award for Translational Bioinformatics
- Clinical Research Informatics (CRI) Distinguished Paper Award
- TBI Student Paper Award
- CRI Student Paper Award

We thank you for your contribution and participation!

Nicholas Tatonetti, PhD
Chair, 2018 TBI Scientific Program Committee
Columbia University

Jyotishman Pathak, PhD
Chair, 2018 CRI Scientific Program Committee
Weill Cornell Medicine
2018 AMIA Board of Directors

Officers

Peter J. Embi, MD, MS, FACMI
Regenstrief Institute
Chair

Thomas H. Payne, MD, FACMI
University of Washington
Past Chair

Titus K. Schleyer, DMD, PhD, FACMI
Regenstrief Institute/Indiana University School of Medicine
Treasurer

Theresa A. Cullen, MD
Secretary

Directors

Tiffany J. Bright, PhD
Agency for Healthcare Research and Quality

Wendy W. Chapman, PhD, FACMI
University of Utah

Sarah A. Collins, RN, PhD
Partners Healthcare System/Brigham and Women’s Hospital/Brigham and Women’s Hospital/Harvard Medical School

David Dorr, MD, MS, FACMI
Oregon Health & Science University

Paul Fu, Jr., MD, MPH, FAAP
Los Angeles County Harbor-UCLA Medical Center/University of California, Los Angeles

Wanda Pratt, PhD, FACMI
University of Washington

Gretchen Purcell Jackson, MD, PhD
Vanderbilt University Medical Center

Neil Sarkar, PhD, FACMI
Brown University

Dean F. Sittig, PhD, FACMI, FHIMSS
University of Texas Health Science Center at Houston
Jessica Tenenbaum, PhD
Duke Translational Medicine Institute

Adam Wright, PhD
Brigham and Women’s Hospital/Harvard Medical School

Ex-Officio Board Members

Christopher G. Chute, MD, DrPH, FACMI
Johns Hopkins University
American College of Medical Informatics President

Douglas B. Fridsma, MD, PhD, FACP, FACMI
AMIA
President and Chief Executive Officer

Stephen Johnson, PhD, FACMI,
Academic Forum Executive Committee Chair

Albert M. Lai, PhD
Washington University
Working Group Steering Committee Representative

Shauna M. Overgaard, MHI, PhD Candidate
University of Minnesota
Student Working Group Representative

Howard R. Strasberg, MD, MS, FACMI
Wolters Kluwer Health
Informatics Partnership Council Chair
2018 CRI Summit Scientific Program Committee

Chair
Jyotishman Pathak, PhD
Weill Cornell Medicine

Vice Chairs
Hua Xu, PhD
The University of Houston Health Science Center at Houston

Genevieve Melton Meaux, MD, PhD
University of Minnesota

Members
Douglas Bell, MD
University of California Los Angeles/David Geffen School of Medicine

Trevor Cohen, MBChB, PhD
University of Texas Health Science Center at Houston

David Dorr, MD, MS, ACMI
Oregon Health & Science University

Marcelline Harris, PhD
University of Michigan

Guoqian Jiang, MD, PhD
Mayo Clinic College of Medicine

Peggy Peissig, PhD
Marshfield Clinic Research Foundation

William Thompson, PhD
Northwestern University

Kavishwar Wagholikar, MD, PhD
Harvard Medical School/Massachusetts General Hospital

Meliha Yetisgen, PhD
University of Washington

Yiye Zhang, PhD, MS
Weill Cornell Medicine/NewYork-Presbyterian
<table>
<thead>
<tr>
<th>AMIA 2018 Informatics Summit Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abend, Aaron H</td>
</tr>
<tr>
<td>Adam, Terrence J</td>
</tr>
<tr>
<td>Afzal, Naveed</td>
</tr>
<tr>
<td>Aguiar, Rute</td>
</tr>
<tr>
<td>Aguiar-Pulido, Vanessa</td>
</tr>
<tr>
<td>Akbilgic, Oguz</td>
</tr>
<tr>
<td>Aldekhyyel, Raniah</td>
</tr>
<tr>
<td>Alekseyenko, Alexander V</td>
</tr>
<tr>
<td>AlHuneiti, Rasmeh</td>
</tr>
<tr>
<td>Alipanah, Neda</td>
</tr>
<tr>
<td>Alonso-Calvo, Raul</td>
</tr>
<tr>
<td>Alqahtani, Khaled Saad</td>
</tr>
<tr>
<td>Aluthge, Dilum Priyanka</td>
</tr>
<tr>
<td>Anand, Edwin</td>
</tr>
<tr>
<td>Anand, Vibha</td>
</tr>
<tr>
<td>Anderson, Nick</td>
</tr>
<tr>
<td>Annechiarico, Robert P</td>
</tr>
<tr>
<td>Aran, Dvир</td>
</tr>
<tr>
<td>Arruda-Olson, Adelaide M.</td>
</tr>
<tr>
<td>Arvanitis, Theodoros N.</td>
</tr>
<tr>
<td>Atnoor, Deven</td>
</tr>
<tr>
<td>Bahroos, Neil</td>
</tr>
<tr>
<td>Balasubramanian, Jeya Balaji</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>Candia, Julian</td>
</tr>
<tr>
<td>Cantor, Michael</td>
</tr>
<tr>
<td>Cato, Kenrick D.</td>
</tr>
<tr>
<td>Chandran, Uma</td>
</tr>
<tr>
<td>Chandras, Rajan</td>
</tr>
<tr>
<td>Chaudhary, Kumardeep</td>
</tr>
<tr>
<td>Chen, Austin</td>
</tr>
<tr>
<td>Chen, Bin</td>
</tr>
<tr>
<td>Chen, David</td>
</tr>
<tr>
<td>Chen, David T.</td>
</tr>
<tr>
<td>Chen, Elizabeth S</td>
</tr>
<tr>
<td>Chen, Guocai</td>
</tr>
<tr>
<td>Chen, Jonathan H</td>
</tr>
<tr>
<td>Chen, Tianlong</td>
</tr>
<tr>
<td>Chen, Yang</td>
</tr>
<tr>
<td>Chen, You</td>
</tr>
<tr>
<td>Cheng, Qiang</td>
</tr>
<tr>
<td>Cheng, Shenghui</td>
</tr>
<tr>
<td>Chi, Chih-Lin</td>
</tr>
<tr>
<td>Chin, Lauren</td>
</tr>
<tr>
<td>Choudhury, Olivia</td>
</tr>
<tr>
<td>Chu, Yuan-Chia</td>
</tr>
<tr>
<td>Chute, Christopher G</td>
</tr>
<tr>
<td>Cimino, James J</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>Falconer, Seanne</td>
</tr>
<tr>
<td>Fan, Jung-wei</td>
</tr>
<tr>
<td>Fan, Yadan</td>
</tr>
<tr>
<td>Fang, Daniel Z</td>
</tr>
<tr>
<td>Farcas, Claudiu</td>
</tr>
<tr>
<td>Faust, Louis</td>
</tr>
<tr>
<td>Fearn, Paul</td>
</tr>
<tr>
<td>Fedorov, Andriy</td>
</tr>
<tr>
<td>Feld, Shara</td>
</tr>
<tr>
<td>Feldman, Keith</td>
</tr>
<tr>
<td>Finkelstein, Joseph</td>
</tr>
<tr>
<td>Fiszman, Marcelo</td>
</tr>
<tr>
<td>Fohner, Alison E</td>
</tr>
<tr>
<td>Fontelo, Paul</td>
</tr>
<tr>
<td>Foraker, Randi E</td>
</tr>
<tr>
<td>Fort, Daniel</td>
</tr>
<tr>
<td>Foster, Marva</td>
</tr>
<tr>
<td>Freimuth, Robert R</td>
</tr>
<tr>
<td>Friedman, Carol</td>
</tr>
<tr>
<td>Fultz Hollis, Kate</td>
</tr>
<tr>
<td>Gabriel, Davera Lea</td>
</tr>
<tr>
<td>Gallego, Blanca</td>
</tr>
<tr>
<td>Garcia, Tzintzuni</td>
</tr>
<tr>
<td>Garcia-Arce, Andres</td>
</tr>
<tr>
<td>Kunz, Isaac</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Kurc, Tahsin</td>
</tr>
<tr>
<td>Kury, Fabricio Sampaio Peres</td>
</tr>
<tr>
<td>Kuusisto, Finn</td>
</tr>
<tr>
<td>Kwon, KiBeom</td>
</tr>
<tr>
<td>Lacson, Ronilda</td>
</tr>
<tr>
<td>Laderas, Ted</td>
</tr>
<tr>
<td>Landis-Lewis, Zach</td>
</tr>
<tr>
<td>Heermann Langford, Laura</td>
</tr>
<tr>
<td>LaPolla, Fred</td>
</tr>
<tr>
<td>Ledbetter, David</td>
</tr>
<tr>
<td>Lee, Chiaju</td>
</tr>
<tr>
<td>Lee, Dae Hyun</td>
</tr>
<tr>
<td>Lee, Jerry Ssu-Hsien</td>
</tr>
<tr>
<td>Lee, Younghee</td>
</tr>
<tr>
<td>Lehmann, Harold P.</td>
</tr>
<tr>
<td>Lenert, Leslie</td>
</tr>
<tr>
<td>LePendu, Paea</td>
</tr>
<tr>
<td>Leung, Tiffany I</td>
</tr>
<tr>
<td>Lever, Jake</td>
</tr>
<tr>
<td>Li, Fuhai</td>
</tr>
<tr>
<td>Li, Haiqing</td>
</tr>
<tr>
<td>Li, Haiquan</td>
</tr>
<tr>
<td>Li, Haoran</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Rasmussen, Luke V</td>
</tr>
<tr>
<td>Ratanawongsa, Neda</td>
</tr>
<tr>
<td>Rath, Sachi</td>
</tr>
<tr>
<td>Read, Kevin B</td>
</tr>
<tr>
<td>Redd PhD, Douglas F</td>
</tr>
<tr>
<td>Rehman, Shakaib U</td>
</tr>
<tr>
<td>Reimer, Andrew P</td>
</tr>
<tr>
<td>Ritchie, Marylyn D</td>
</tr>
<tr>
<td>Riva, Alberto</td>
</tr>
<tr>
<td>Rizvi, Rubina</td>
</tr>
<tr>
<td>Roberts, Kirk</td>
</tr>
<tr>
<td>Rodriguez, Laritza</td>
</tr>
<tr>
<td>Romano, Joseph</td>
</tr>
<tr>
<td>Ronquillo, Jay G</td>
</tr>
<tr>
<td>Rooney, Robert J</td>
</tr>
<tr>
<td>Rosenman, Marc</td>
</tr>
<tr>
<td>Rossi, Lorenzo</td>
</tr>
<tr>
<td>Rostami, Reihaneh</td>
</tr>
<tr>
<td>Rotenberg, David</td>
</tr>
<tr>
<td>Rouhizadeh, Masoud</td>
</tr>
<tr>
<td>Ru, Boshu</td>
</tr>
<tr>
<td>Rusincovitch, Shelley A</td>
</tr>
<tr>
<td>Ryan, Nathan</td>
</tr>
<tr>
<td>Ryu, Euijung</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td>Stallings, Sarah C</td>
</tr>
<tr>
<td>Starren, Justin</td>
</tr>
<tr>
<td>Steinbach, Michael</td>
</tr>
<tr>
<td>Stephan, Christina L</td>
</tr>
<tr>
<td>Stewart, Walter F.</td>
</tr>
<tr>
<td>Strasberg, Howard R</td>
</tr>
<tr>
<td>Strauss, John</td>
</tr>
<tr>
<td>Stroup, Richard E</td>
</tr>
<tr>
<td>Su, Wu-Chen</td>
</tr>
<tr>
<td>Subbian, Vignesh</td>
</tr>
<tr>
<td>sun, deyu</td>
</tr>
<tr>
<td>Sun, Jimeng</td>
</tr>
<tr>
<td>Sun, Zhaonan</td>
</tr>
<tr>
<td>Sura, Andrei</td>
</tr>
<tr>
<td>Surkis, Alisa</td>
</tr>
<tr>
<td>Sward, Katherine A</td>
</tr>
<tr>
<td>Syed-Abdul, Shabbir</td>
</tr>
<tr>
<td>Tajgardoorn, Mohammad</td>
</tr>
<tr>
<td>Tao, Cui</td>
</tr>
<tr>
<td>Tao, Shiqiang</td>
</tr>
<tr>
<td>Tawailo-Graovac, Maja</td>
</tr>
<tr>
<td>Taweel, Adel</td>
</tr>
<tr>
<td>Overby Taylor, Casey</td>
</tr>
<tr>
<td>Teal, Evgenia</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>---------------------------</td>
</tr>
<tr>
<td>West, Vivian L</td>
</tr>
<tr>
<td>Westra, Bonnie L</td>
</tr>
<tr>
<td>Wiegreffe, Sarah</td>
</tr>
<tr>
<td>Wieland, Daryl L</td>
</tr>
<tr>
<td>Wilcox, Adam</td>
</tr>
<tr>
<td>Wiley, Laura K</td>
</tr>
<tr>
<td>Wiley, Jr., Ken</td>
</tr>
<tr>
<td>Willett, DuWayne Lee</td>
</tr>
<tr>
<td>Williams, Marc S</td>
</tr>
<tr>
<td>Wimbiscus, Laura</td>
</tr>
<tr>
<td>Woodcock, Deborah</td>
</tr>
<tr>
<td>Wood-Wentz, Christina</td>
</tr>
<tr>
<td>Wroblewska, Aleksandra</td>
</tr>
<tr>
<td>Wu, Danny</td>
</tr>
<tr>
<td>Wu, Xi</td>
</tr>
<tr>
<td>Wu, Yonghui</td>
</tr>
<tr>
<td>Wyatt, Matthew C</td>
</tr>
<tr>
<td>Xiao, Ran</td>
</tr>
<tr>
<td>Xirasagar, Sandhya</td>
</tr>
<tr>
<td>Xu, Julia</td>
</tr>
<tr>
<td>Xu, Shen</td>
</tr>
<tr>
<td>Xu, Yifan</td>
</tr>
<tr>
<td>Yahi, Alexandre</td>
</tr>
<tr>
<td>Yan, Xiaowei Sherry</td>
</tr>
</tbody>
</table>
Developing data management infrastructure supporting innovation in multicenter clinical trials at-scale: a report from the Trial Innovation Network

Davera Gabriel, RN\textsuperscript{1}; Bernard LaSalle, BS\textsuperscript{4}; Brian McCourt, BS\textsuperscript{1}; Paul Harris, PhD\textsuperscript{3}
\textsuperscript{1}Duke Clinical Research Institute, Durham, NC; \textsuperscript{2}University of Utah, Salt Lake City, UT; \textsuperscript{3}Vanderbilt University, Nashville, TN;

Abstract

The Trial Innovation Network (TIN) is a collaborative, national multidisciplinary nexus that focuses on improvements in operations, quality and innovation utilizing cooperative learning approaches to support multi-center clinical trials, leveraging the resources and expertise of awardees of the National Center for Advancing Translational Science (NCATS) Clinical and Translational Science Awards (CTSAs). The panel consists of representatives from the three Trial Innovation Centers (TICs) Data Management and Harmonization Working Group (DMHWG) and the Recruitment Innovation Center (RIC). Data management innovation and activities undertaken by the group are discussed.

Introduction

The Trial Innovation Network (TIN) is a collaborative, national multidisciplinary nexus that focuses on improvements in operations, and quality and innovation utilizing cooperative learning approaches to support multi-center clinical trials, leveraging the resources and expertise of awardees of the National Center for Advancing Translational Science (NCATS) Clinical and Translational Science Awards (CTSAs).\textsuperscript{1} A key goal of the TIN is to “increase the efficiency and effectiveness of clinical trials and studies by focusing on innovation, excellence, and collaboration; leveraging the strength and expertise of the CTSA Program; and fostering partnerships between the Institutes and Centers of the National Institutes of Health (NIH), researchers, participants, providers, industry, and other stakeholders.”\textsuperscript{2} The network is comprised of a RIC led by Vanderbilt University and three TICs based at Duke Clinical Research Institute, Johns Hopkins University/Tufts University and the University of Utah. The TIN partners: the RIC, TICs and CTSA hubs, collaborate to advance deployment of operational and technical infrastructure necessary to support multicenter clinical trials as requested by clinical investigators, while developing, adopting, and evaluating innovative approaches to recruitment, operational practices, deployment methods, and utilization of information technology. All participants in this panel have agreed to take part in this presentation.

TIN Data Management and Harmonization Working Group

Data collection and management are among the challenges and key drivers of study costs for clinical trials and as such present opportunities for innovative adoption of best practices and technology advancement. Investigators and sponsors often spend considerable time and resources building new research infrastructure for individual studies, or lack timely progress due to unmet data and information technology requirements. Standardized approaches to data collection and management could accelerate study startup time and reduce costs because studies would not have to start totally \textit{de novo} each time. The mission of the TIN Data Management and Harmonization Working Group is to offer data management services to TIN applicants that support innovative trial designs, streamlined data collection, utilize harmonized data standards, and coordinated processes for study monitoring and safety reporting.\textsuperscript{2}

Each institution shared strategies and lessons learned with other working group members in order to develop a program-level, harmonized approach to “ways of doing the work of clinical trials…embedding innovation and critical evaluation into the process.”\textsuperscript{2} This collaboration approach as well as the work products serve as an example to the informatics community of the value of collaboration among independent clinical research organizations and a model for similar collaborative research data management programs. Panel presentations will showcase:

Harmonized Data Management Services and Metrics

One of the goals of the DMHWG was to identify Clinical Data Management (CDM) services delivered by each TIC and harmonize the associated scope, inputs, and deliverables and operational metrics. Principal Investigators (PIs) need to have consistent expectations of the data management practices independent of which Data Center is providing support. Adopting consistent definitions, the TICs are then able to achieve and measure service delivery without being prescriptive about the technology, platforms or procedures utilized by any individual TIC.
**Approach to Data Standards Adoption**

The DMHWG has advanced vanguard efforts toward NCATS vision of interoperable data for conduct of multi-center clinical trials. These efforts include identification of a common set of coded data elements (CDEs) for data collection as well as safety and Data and Safety Monitoring Board (DSMB) use cases. The DMHWG collaborated by sharing data collection practices, evaluated existing resources and common requirements to support creating interoperable data and selected Clinical Data Interchange Standards Consortium (CDISC) standards: Clinical Data Acquisition Standards Harmonization (CDASH) terminology and Study Data Tabulation Model (SDTM) data elements and therapeutic area standard data elements as core data standards. Additionally, the group has developed a harmonized workflow for utilization of CDE’s and submission of editorial requests when data required for a trial is not yet available in a published standard. This presentation will describe the processes used to determine the scope and priorities of the data standards adoption project.

**Innovations utilizing the REDCap® platform**

Vanderbilt University in partnership with Duke University is leading TIN data management activities to advance innovative utilization of Research Electronic Data Capture (REDCap) and other technologies that streamline enrollment and data collection processes and increase trial data quality. Advanced features of REDCap, including the Dynamic Data Pull (DDP), have been built into data management plans for one TIN study in order to speed data collection, increase the consistency and quality of data and effect reduction of costs. Electronic consent (eConsent) is an innovative service being offered to TIN trial data management recipients via REDCap that could create cost savings, and increase participant knowledge and satisfaction. REDCap has features that streamline the creation of study-specific databases or registries.

**Safety Reporting**

The DHMWG utilized its open and collaborative approach to identify and assess the Adverse Event (AE) systems and processes utilized by the TICs. The group is defining a set of safety surveillance standards, which includes harmonizing AE reporting, remote monitoring processes, standardizing safety components of DSMB reporting templates, standard statistical output, reporting templates for adverse events and other clinical data monitoring activities that lend themselves to scalable robust quality control. The group is reviewing the CDISC Adverse Event Domain for common data elements supporting safety requirements and guidance from CDASH Serious Adverse Event (SAE) Supplement in order to utilize defined, interoperable data elements for Individual Case Safety Reports (ICSR) to regulatory authorities.

**Conclusion**

In summary, we have created a cross-institution network across the RIC and the three TICs with shared interest in advancing the goals of the TIN: to develop a collaborative approach to common data management services required to support multi-center clinical trials. The network has improved CDM by harmonizing processes, strategies and leveraging accessible informatics methods and tools. The Data Management and Harmonization Working Group aims to provide measurement data on important data processing, collection and data management processes with the goal of supporting a common set of data management best practices in order to provide a consistent platform to promulgate innovative data management approaches, tools and infrastructure supporting the Network.

**Acknowledgements**

The work presented in this panel reflects the collaboration of teams of individuals from across TIN member organizations. The panelists would like to acknowledge the contributions of Steve Mayo and Rachel Dlugash representing Johns Hopkins University; Ken Gersing and Tiina Urv of NCATS; Kathy Sward and Jeff Yearley of the University of Utah; Stephany Duda of Vanderbilt University, and the entire TIC Data Management and Harmonization Working Group. Portions of this work have been funded through the National Center for Advancing Translational
Sciences, National Institutes of Health, under award numbers U24TR001608, U24TR001597, U24TR001609, and U24TR001579.

Questions to enhance audience participation

1) What are tradeoffs between harmonization/standardization and innovation?
2) What leads to / how can we overcome investigator resistance to using standardized data collection instruments?
3) Why are metrics a critical aspect of network innovation?
4) What barriers would you face in order to implement a harmonized approach to safety reporting at your organization?
5) Where do you see eConsent on the technology adoption curve? Is it gaining acceptance at your organization?

References

The Future of Evidence Synthesis: Incorporating heterogeneous and patient-specific evidence towards Evidence-based Precision Medicine

Ralph Horwitz, MD, MACP 1, John Ioannidis, MD, DSc2, Jonathan Silverstein, MD, MS, FACS, FACMI3, Ida Sim, PhD, MD4, Jodi Schneider, MA, MSLIS, PhD5
1 Temple Translational Medicine Institute, Lewis Katz School of Medicine, Temple University, Philadelphia, PA; 2School of Medicine, Stanford University, Stanford, CA; 3Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA; 4School of Medicine University of California, San Francisco, San Francisco, CA; 5School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL

Abstract As clinical practice becomes more focused on personalized and precision medicine, there are increasing challenges for evidence-based practice, which combines the best research evidence, clinical expertise, and patient values and preferences in order to support clinical decision making. Identifying the best research evidence relevant to a particular patient requires new approaches to evidence synthesis. New technologies are emerging for both the current standard (group-oriented Randomized Controlled Trials) and “beyond the RCT” studies.

Timeliness of the Topic

Recently there has been an explosion of new technologies for assessing evidence, resulting in a vigorous and fast-moving field at the intersection of evidence-based medicine and precision medicine. Evidence-based precision medicine investigates a key question: how to generate high-quality assessments of the evidence that are customized to specific patients and produced fast enough to be relevant to patient care right now. Current approaches (such as writing systematic reviews) are too slow and labor-intensive, and do not always provide actionable recommendations useful for clinicians dealing with individual patients, or even for clinical guidelines.

General Description of the Panel

Innovation brings new opportunities, but also brings further challenges to the interpretability and comparability of data for evidence syntheses. Using patient-level data and real-world evidence for evidence-based medicine will require a better understanding of the comparability and relative strengths/disadvantages between these “beyond the RCT” studies and the current standard: group-oriented Randomized Controlled Trials. The panel will cover emerging trends in clinical evidence gathering as well as how technology is enabling novel approaches to evidence synthesis tasks that inform precision medicine.

Informaticists and methodologists have been working to co-develop new technologies and new methods for evidence-based precision medicine. Such innovations already have a large impact, including:
- **Synthesizing evidence more quickly.** Partial automation of certain tasks in systematic reviewing has become feasible. Quicker evidence synthesis, especially for scoping and rapid reviews, is enabled by technologies such as text mining, machine learning, and ontologies. Living reviews can continually update a given research synthesis.
- **Maximizing the use of clinical trial data.** Prospective meta-analyses, network meta-analyses, use of individual-level data, and umbrella reviews, leverage existing clinical trial data.
- **Generating evidence with novel trial design.** In technology assisted trials of Parkinson’s medication, researchers at SAGE, Pfizer, and IBM are using mobile phones and biosensors to monitor enormous numbers of patients, pinpointing whether and when medication is working for one specific patient.
- **Deriving best practices from real-world health records.** Learning health systems are becoming reality. Real-world health records can be more quickly aggregated and synthesized. Readmission risk, for example, is being widely studied due to health system incentives.
Panelists

Ida Sim, MD, PhD is a primary care physician, informatics researcher, and entrepreneur. She is a Professor of Medicine at the University of California, San Francisco, where she co-directs Biomedical Informatics at UCSF’s Clinical and Translational Sciences Institute. She is Consortium Core Lead with the Mobile Data to Knowledge NIH Center of Excellence. Her current research focuses on the use of mobile apps and sensors to improve health and manage disease for populations and individuals, and to make clinical research faster and less expensive.

John Ioannidis, MD, DSc is C.F. Rehnborg Chair in Disease Prevention at Stanford University, Professor of Medicine, Professor of Health Research and Policy, Professor (by courtesy) of Biomedical Data Science, Professor (by courtesy) of Statistics; co-Director, Meta-Research Innovation Center at Stanford; and Director of the PhD program in Epidemiology and Clinical Research.

Ralph I. Horwitz, MD, MACP, is Professor of Medicine and Director, Transformative Medicine Institute, Temple University and Harold H. Hines, Jr. Professor Emeritus of Medicine and Epidemiology at Yale University. His work advances the field of population health research and expands translational research into factors and interventions that influence the health of individuals and populations.

Jonathan C. Silverstein, MD, MS, FACS, FACMI serves as Chief Research Informatics Officer and Professor of Biomedical Informatics at University of Pittsburgh School of Medicine. He is internationally known for the application of advanced computing architectures to biomedicine and on the design, implementation and evaluation of high-performance collaboration and visualization environments for anatomic education and surgery.

Moderator

Jodi Schneider, MA, MSLIS, PhD is Assistant Professor of Information Sciences at the University of Illinois at Urbana-Champaign. She is a researcher in biomedical knowledge management, focusing on biomedical ontologies, evidence synthesis methodologies, and systematic review automation. She is a co-investigator on the NIH-funded Evidence Based Medicine project headed by Aaron Cohen and Neil Smalheiser which focuses on development of automated tools to assist in the writing of systematic reviews.

Presentations

In this 90-minute panel, we plan four brief presentations followed by an extensive Q&A session.

Dr. Sim will open the panel with an overview of the emerging trends in evidence-based medicine. She will talk about the emerging availability of participant-level clinical trial data\(^1\) and what that means for evidence synthesis\(^9\) such as living systematic reviews. She will also describe new tech-enabled study designs that are enabling more personalized evidence generation.

Dr. Ioannidis will then discuss the present and future of specific types of evidence-based medicine analyses including network meta-analyses: prospective meta-analysis, individual-level data, networks and umbrella reviews.\(^6,7\)

Dr. Horwitz will discuss evidence generation for individualized decision making, focused on addressing the question “Will a given therapeutic regimen help my patient at a given point in her/his clinical course?”\(^4,5\)

Dr. Silverstein will introduce and use the learning health systems framework\(^3\) to offer a very applications-focused talk about the secondary use of electronic health records\(^8\) around the full loop of innovations, with examples from clinical/public health, genomics, and sensors.\(^2\) He will describe the potential for personalized analytics: how artificial intelligence can cluster patients into fine-grained subgroups to determine the personalized treatment that may be best for one particular patient.

Finally, Dr. Schneider will lead the Q&A, and facilitate a broad, wide-ranging discussion on the major theme of the panel, which is how future evidence synthesis can support evidence-based precision medicine, by incorporating heterogeneous and patient-specific evidence.
Anticipated Audience

The primary audience for this panel will include clinical research informaticians, clinical informaticians, evidence-based medicine researchers, and clinical trial practitioners. The panel will also be of interest to health information and knowledge management industry professionals related to clinical research and translational research and designers and developers of Electronic Health Records and patient registries.

Discussion Questions to Enhance Audience Participation

- What are the major roadblocks to progress towards evidence-based precision medicine?
- What are the first actionable steps along the way?
- What are the major take home messages do you have for CMIOs and clinicians? For researchers in clinical informatics and evidence-based medicine?

Learning Objectives

- Participants will be able to describe evidence-based medicine and evidence synthesis in their own words.
- Participants will demonstrate awareness of emerging trends in clinical evidence gathering and how this could potentially impact evidence-based medicine.
- Participants will demonstrate awareness of how technology is enabling novel approaches to evidence synthesis tasks that inform precision medicine.

Statement

All listed persons have participated in the creation of this proposal and have agreed to participate in the panel presentation.

Acknowledgements

JS was supported by National Library of Medicine grant R01LM010817: "Text Mining Pipeline to Accelerate Systematic Reviews in Evidence-based Medicine".

References

Improving Standardization and Interoperability of Common Data Models for Clinical and Translational Research

Guoqian Jiang, MD, PhD¹, Jon Duke, MD², Daniella Meeker, PhD³, Eliel Oliveira⁴, MS, MBA, Mitra Rocca⁵, Harold Solbrig, MS¹, Shawn Murphy, MD, PhD⁶

¹Mayo Clinic, Rochester, MN; ²Georgia Institute of Technology, Atlanta, GA; ³University of Southern California, Los Angeles, CA; ⁴REACHnet, New Orleans, LA; ⁵FDA, Silver Spring, MD; ⁶Harvard Medical School, Boston, MA

Abstract

Clinical and translational research studies increasingly involve the manipulation of large datasets. The development of integrated data repositories (IDRs) based on common data models (CDMs) are needed to both lower the effort required for, as well as incentivize the process of, standardizing clinical research data in clinical and translational research. In particular, the harmonization of CDMs can make it easier for researchers to share their data and research results across different informatics platforms and to broadly benefit the community. This thematic panel will highlight five projects that are developing technology infrastructure and tools for the CDM harmonization, standardization and applications. Attendees will learn about novel open source CDM-based informatics platform, and how CDMs are used in data sharing platforms. Attendees can engage with panelists on the CDM harmonization and interoperability challenges and potential solutions in support of clinical and translational research applications.

Introduction

Data integration has become a prerequisite for conducting clinical and translational research. As the number of datatypes and sources continues to grow, the development of the integrated data repositories (IDRs) based on common data models (CDMs) are needed to both lower the effort required for, as well as incentivize the process of standardizing clinical research data. This is especially important for clinical research investigators with limited computing resources. In a 2010 survey conducted by the Clinical and Translational Science Award (CTSA) consortium, IDR was defined as a data warehouse integrating various sources of clinical data to support queries for a range of research-like functions. Survey results suggest that individual organizations are progressing in their approaches to the development, management, and use of IDRs as a means to support a broad array of research. A variety of data models have been developed to provide a standardized data interface that supports organizing clinical research data into a standard structure in such IDRs. These data models include the Observational Medical Outcomes Partnership (OMOP) CDM, the National Patient-Centered Research Networks (PCORnet) CDM, the Informatics for Integrating Biology and the Bedside (i2b2) Star Schema and the HL7 Fast Healthcare Interoperability Resources (FHIR).

The ultimate goal of a method to harmonize CDMs is to identify patient cohorts or results based on a single query that can be used across multiple implementations. Typically, CDMs are deployed across independent organizations and either the query or data are harmonized so that a consolidated result can be produced. This thematic panel will highlight five projects that are developing technology infrastructure and tools for the CDM harmonization, standardization and applications. The use cases and application domains include the integration of OHDSI CDM with FHIR (e.g., the OMOP on FHIR project and the HL7 DAF project), the CDM Harmonization for drug safety surveillance, the integration of i2b2 with FHIR (e.g, FHIR in i2b2 project), and the i2b2 on OMOP project. Notably, an OHDSI FHIR Workgroup has been established to enable a community-based consensus development effort on the harmonization of the OHDSI CDM and FHIR, whereas an i2b2 FHIR Workgroup is under planning.

Topics Relevant to CDM Harmonization and Interoperability

In this panel, we highlight five CDM harmonization projects that describe various platforms to demonstrate the informatics platform and framework applying metadata that would drive data discovery in the era of rapidly accumulating big data. These initiatives are described briefly below:
1. **OMOP on FHIR (Jon Duke)**

The Observational Medical Outcomes Project (OMOP) model is a common data model that has been adopted globally for a wide range of research activities. The OMOP model is accompanied by a well-maintained vocabulary that connects over 50 ontologies and establishes standard concepts for each clinical domain. Many analytics tools have been built around OMOP, including the OHDSI software stack which designed for standardized population-level and patient-level analyses. The OMOP model is highly complementary to FHIR, a rapidly growing health information exchange standard that supports API-based exchange of health information content in a standard simplified data model to facilitate information sharing between health information systems. In this section of the panel, we discuss the open-source OMOP-on-FHIR project, which provides automated transformation bi-directionally between OMOP and FHIR and has been used for numerous projects ranging from disease surveillance, clinical decision support, and drug safety.

2. **HL7 Data Access Framework (DAF) (Daniella Meeker, Eliel Oliveira)**

The DAF-Research initiative developed and implemented FHIR standards for enabling access to data from multiple organizations in the context of the Learning Health System. DAF, initiated in 2013, is focused on the identification and standards necessary to access and extract US Core Data Elements from within an organization’s health IT systems (DAF 1), from an external organization’s health IT systems (DAF 2), or from health IT systems across multiple organizations (DAF-Research). DAF 1 and DAF 2 were focused on retrieving individual-level data. The same requirements standards, and interoperability are required for research and population health. While DAF-Research was conceived originally as a way to support research networks like PCORnet, DAF-Research inherently includes capabilities that are devoted to populations of individuals. DAF addresses not only research needs, but also population health and business intelligence needs within or across organizations. Therefore, Value Based Care and Precision Medicine (Sync for Science) stakeholders are key beneficiaries of successful DAF implementations. Two research networks participated in the DAF pilots to create implementation guides. One network (REACHNet) used PCORnet CDM as a foundation for creating FHIR-based Application Programming Interfaces; the other used OMOP (pSCANNER), creating resources that can be collectively reused for dozens of health systems and tens of millions of patients. We will discuss how randomized trial eligibility criteria and the Clinical Quality Framework can leverage all components of DAF to promote a Learning Health System. We show how DAF can produce more rapid translation between technical specifications for research and computable guidelines and decision support.

3. **FDA Harmonization of various Common Data Models and Open Standards for Evidence Generation (Mitra Rocca)**

In order to achieve a sustainable data network infrastructure, promote interoperability and foster the creation of a Learning Health System (LHS) as laid out in the Connecting Health and Care for the Nation a Shared Nationwide Interoperability Roadmap, there is a need to map and transform data across various Common Data Models (CDMs) and leverage open-source standards. This project is aimed at facilitating the use of Real World Data (RWD) sources (e.g., claims, Electronic Health Records (EHRs), registries, electronic Patient Reported Outcomes (ePRO)) to support evidence generation for regulatory and clinical decision-making. This project is aimed at advancing the maturity and sophistication of the Office of the Secretary’s Patient-Centered Outcomes Research (PCOR) Trust Fund (OS-PCORTF) functionalities and the national data capacity for conducting PCOR studies. In this project, FDA collaborates closely with the National Institutes of Health (NIH) and the Office of the National Coordinator for Health Information Technology (ONC).

4. **FHIR in i2b2 (Harold Solbrig, Guoqian Jiang)**

The observation fact table lies at the heart of the i2b2 data repository. Patient information is recorded in this table as sets of atomic "facts", each of which is associated with a concept identifier. Concept identifiers are, in turn, defined and organized in the i2b2 ontology. The ability to manipulate both the observation fact concepts and their arrangement in the ontology provides researchers with a wealth of flexibility. It is this same flexibility, however, that interferes with the dissemination and integration of research data, and has led to calls for shared i2b2 ontologies. One promising approach to this has been to leverage the emerging capabilities of HL7 FHIR, the emerging next generation standards framework for the exchange of electronic health records (EHR)-data. The latest version of HL7 FHIR (STU3) includes a proposed standard for the representation of FHIR resources in RDF, a component of which is the "FHIR Metadata Vocabulary (FMV)" -- definitions and organizational information for the URI's (identifiers) that appear in the individual FHIR resources. We have
developed a set of ETL scripts that allow FHIR-based data to be loaded into the i2b2 observation fact tables using the FHIR RDF URI's as the concept identifiers, with associated patient, provider and encounter(visit) information is being loaded into the corresponding i2b2 supporting infrastructure. We then loaded the FMV itself as part of the i2b2 ontology and have used it to demonstrate that meaningful research questions can be answered using this infrastructure. We discuss the role of the FHIR RDF technology on facilitating data transformation and quality assurance, and the challenges in integrating HL7 FHIR into i2b2, including handling terminological value sets bound to the FHIR data models.

5. **I2b2 on OMOP (Shawn Murphy)**

The data models of i2b2 and OMOP have many similarities. We take advantage of these similarities to offer an evolution of the i2b2 software that adapts to the OMOP data model. This allows the query formulation in the i2b2 software that relies on the i2b2 Application Programming Interface (API) to be utilized on top of an OMOP data source. As a result, most of the functionality of i2b2 Software is preserved on an OMOP data source (with proper respect to the OHDSI ontologies). As a result a SHRINE tool or a SMART-on-FHIR tool that relies on the i2b2 API, over OHDSI ontologies, can run with an OMOP data source.

This panel is timely and relevant given the urgent need to integrate clinical research data to facilitate clinical and translational data discovery and analytics, and advance our understanding of human health and disease. Initiatives such as the precision medicine and cancer moonshot will spur the generation of large datasets. We need appropriate tools and technologies applying CDMs to manage and analyze these to inform clinical research and patient care.

**Discussion questions**

1. What would you like to see in a CDM-based research-data sharing platform?
2. What types of metadata/data should be shared?
3. What CDM-based tools will be most useful to help your translational research?
4. What are your most critical metadata/data sharing challenges?

**Learning Objectives**

1. Attendees will learn about novel open source CDM-based informatics platform
2. Attendees will learn about how CDMs are used in data sharing platforms
3. Attendees can engage with panelists on CDM harmonization challenges and potential solutions in support of clinical and translational research applications

**Organizer statement:** All speakers have agreed to attend AMIA CRI 2018 Annual Symposium and participate on this panel.
Nursing Documentation and the Clinical Research Informatics Pipeline
Jeffrey G. Klann, PhD\(^1\), Sarah A Collins, RN, PhD\(^2,3\), Kenrick D. Cato, PhD, RN\(^6,7\), Lemuel R. Waitman\(^4\), Bonnie L. Westra, PhD, RN, FAAN, FACMI\(^5\)

\(^1\)Massachusetts General Hospital, Boston, MA; \(^2\)Brigham and Women’s Hospital, Boston, MA; \(^3\)Harvard Medical School, Boston, MA; \(^4\)Center for Medical Informatics and Enterprise Analytics, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, Kansas; \(^5\)University of Minnesota, School of Nursing, Minneapolis, MN, USA; \(^6\)Columbia University School of Nursing, New York, NY; \(^7\)New York Presbyterian Hospital, New York, NY

Abstract
Data research repositories and networks, the backbone of many retrospective informatics research projects, contain selected “most relevant” data transformed from the electronic health record. Data models commonly include patient demographic information, encounter context information, and key clinical data (e.g., medication, diagnosis, and procedure). Nursing data, however, is seldom included. Common Data Models (e.g., OMOP and PCORnet CDMs), research networks (PCORnet and NIH ACT), and ontology/terminology curation organizations (e.g., BioPortal) have largely ignored nursing data. However, in hospitals, nurses provide most direct care to patients, diligently document patient states, and routinely make decisions and implement interventions based on these clinical assessments. Flowsheet data provide structured documentation that goes beyond physiologic measures to assessments of well-being and wellness (e.g., activities of daily living), levels of risk, and responses to treatment. Nursing notes offer additional context, and they can be used to predict morbidity and mortality. [1]

Here, we explore several exciting endeavors at several health systems to bring nursing documentation into the pipeline of clinical research informatics, and how it is being applied to healthcare problems. In this time of heavy investment into data research networks, this panel addresses the timely need to expand our view of what constitutes core clinical data.

Detailed Panel Description
CONCERN Study
Setting the stage for a discussion of Partners’ and Columbia’s efforts is a multi-site research study entitled “Communicating Narrative Concerns Entered by RNs (CONCERN): CDS Communication for Risky Patient States”. CONCERN is performing analytics of EHR documentation patterns to identify risky patient states across the Partners Healthcare System and New York Presbyterian Healthcare System. Later phases of the study will develop and test a SmartApp CDS notification system based on predictive patterns at both study sites. As a multi-site study across two large health systems that use different vendor EHRs, confirmation of shared data definitions and concepts is necessary. While existing standards can be leveraged for many types of data (e.g., labs, medications, diagnoses), standardized definitions for clinical observations and care transitions that occur within a hospital encounter are lacking, such as transfer to/from a clinical unit, off-unit procedures versus on-unit procedures, and unanticipated transfer to intensive care unit. Definitions and information models to standardize the concepts used across the two study sites will be shared. Specific case examples of variability in available data from each site’s EHR and processes used to define equivalent proxy measures will be used illustrate complexities in this multi-site work. Finally, we will identify the nursing note types identified as useful for the aims of this project, including notes generated from semi-structured data (e.g., plan of care notes) and narrative nursing data (e.g., progress notes, free text comments) and the analytical implications of those structures.

Partners Healthcare
Partners Healthcare has a robust clinical data warehousing infrastructure, with an enterprise-wide data warehouse that aggregates data from multiple hospitals. Researchers can request extracts of data from the repository or their own warehouses (utilizing the “i2b2” software) containing a subset of the enterprise warehouse. This is widely used.[2] However, the warehouse does not contain any nursing documentation. Here we discuss our efforts to create a data warehouse of nursing documentation that is compatible with Partners’ warehouse. With a focus on the inpatient setting, we pull structured flowsheet data from EpicCare and organize it into an i2b2 ontology for easy data search and retrieval. We will discuss overall processes and challenges to identify, extract, and process relevant flowsheets and nursing notes from inpatient encounters. Additionally, we are in the process of extracting additional facts from clinical notes and adding them to the data warehouse. We use an NLP tool called mTerms, and we are adapting this
to nursing notes. [3] The output of this process will be the first step in several larger projects to analyze nursing documentation patterns.

**Columbia University**

The Columbia University Medical Center Clinical Data Warehouse (CDW) was established in 1994 and contains longitudinal medical records for over 5 million patients seen at the NewYork Presbyterian Hospital. The CDW contains the majority of inpatient electronic data from Allscripts SCM and our Laboratory Information System, including most nursing notes. Internally, semantic mapping is achieved by the Medical Entities Dictionary (MED), which incorporates other ontologies like SNOMED and UMLS. As part of a multi-site research project with Partners Healthcare we are synchronizing our i2b2 ontology with theirs. Here we discuss our efforts to extract data from our CDW into our i2b2 instance as well as compare our implementation experience to those of the Partners Healthcare team. The i2b2 infrastructure will be used for data mining and NLP to analyze flowsheet, notes, and outcomes data. Additionally, we plan to utilize our data to build a number of generalizable models, including one that describes intensity of nursing work.

**University of Kansas Medical Center**

Since 2011, KUMC has incorporated nursing and interdisciplinary flowsheets within HERON, their i2b2-based data repository. [4] These data (currently 1.39 billion facts in the 2.8 billion fact repository) have provided essential findings for biomedical research studies and prospective trial screening, specifically advancing our ability to support allied health research in rehabilitation science and nutrition. More recently, we have de-identified free-text nursing findings and multidisciplinary notes. We will provide analysis regarding the use of these data over five years for both feasibility queries and analytic sets generated for observational studies. We will discuss pilot observational studies underway and informatics initiatives within the Greater Plains Collaborative and PCORnet to conduct multicenter research using clinical findings derived from multidisciplinary flowsheets. [5]

**University of Minnesota**

The University of Minnesota participates in PCORI and CTSA initiatives. The common data models supporting these initiatives includes “low hanging fruit” of standardized data. Missing is interprofessional assessments and interventions that influence patient outcomes. A pilot study was conducted to model 10 clinical topics from flowsheet data in i2b2, but technical challenges were incurred including: development of an ontology to organize concepts, mapping multiple flowsheet rows to the same concept i.e. pain rating 0-10, cleaning value sets, and the volume of data.[6,7] A subsequent approach was used to increase the generalizability of data in one of the clinical topics: pain management. Flowsheet metadata were extracted from 10 organizations and mapped to a pain information model using a secure internet-based software program – FloMap. A data-driven consensus process resulted in modifying the initial Pain IM concepts and value sets. The idea is to now imbed this information model into i2b2 or other CDMs to expand data for evaluating the effectiveness of pain management strategies.

**Panelists**

Dr. Klann is a data-mining and data interoperability expert who has worked with the core i2b2 and data warehousing team for five years. He presently is an informatics researcher at the Massachusetts General Laboratory of Computer Science, an Instructor of Medicine at Harvard Medical School, and a Director of the Harvard-based PCORnet Clinical Data Research Network (CDRN).

Dr. Cato is a Nurse Researcher/Assistant Professor for NewYork-Presbyterian Hospital and Columbia University School of Nursing, respectively. Dr. Cato’s program of research focuses on the use of data science to investigate ways of improving patient safety, quality of care, and individual health. His current projects include automated data mining of electronic patient records to discover characters about a patient that are often missed by clinicians and development of predictive models for healthcare associated infections.

Dr. Waitman directs the Center for Medical Informatics and Enterprise Analytics at the University of Kansas Medical Center and is Principal Director for the PCORnet Greater Plains Collaborative Clinical Data Research Network. KUMC and his network have advanced the integration of i2b2 and REDCap for not just feasibility determination but data integration and delivery into analytic sets and the PCORnet Common Data Model. His experience prior to KUMC included developing, managing and commercializing inpatient clinical systems with a focus on computerized physician order entry with advanced decision support.

Dr. Westra is an Associate Professor and Director for the Center for Nursing Informatics at the University of Minnesota, School of Nursing. She is experienced in designing and implementing EHRs for community-based practice and informatics research focused on predictive modeling with EHR data from large health systems.
Dr. Collins (moderator) is a Senior Clinical and Nurse Informatician at Brigham and Women’s Hospital, Instructor of Medicine at Harvard Medical School, and Principal Investigator of the multi-site CONCERN research study. She is an experienced critical care nurse and expert in EHR inpatient documentation and communication, with a focus on nursing data patterns.

All panelists named above have agreed to participate. This work was sponsored by National Institute for Nursing Research 1R01NR016941-01, the CRICO Risk Management Foundation of the Harvard Medical Institutions Incorporated, NIH NCATS UL1TR000001 and 1UL1RR033183, and PCORI contract CDRN-1306-04631.

Discussion Questions

- What are some specific data elements in nursing flowsheet data that add richness to the medical record?
- What are the major challenges in integrating nursing flowsheet data into data repositories?
- How are nursing notes similar or different from other clinical narratives?
- In what ways could nursing documentation be important for clinical research informatics?
- How do clinical workflows and policies impact interpretation of structured and unstructured nursing data?

References


Governance of the Pluripotent Research Database

Harold P. Lehmann, MD PhD FACMI¹, William Hersh, MD², Jihad Obeid, MD³, Megan K. Singleton, JD MBE CIP¹, Anthony E. Solomonides, PhD⁴, Umberto Tachinardi, MD PhD FACMI⁵

¹Johns Hopkins University School of Medicine, Baltimore MD; ²Oregon Health & Science University, Portland OR; ³Medical University of South Carolina, Charleston SC; ⁴NorthShore University HealthSystem, Evanston IL; ⁵University of Wisconsin, Madison WI

Abstract

Academic institutions are building large extracts of their electronic health record (EHR) for discoveries, data sharing, registries, multiple machine learning (ML), and other research projects. Many institutional review boards (IRB) are unclear how to provide governance for such pluripotent databases. IRBs are conflicted between the need to protect patients and conform to regulations, through release of the “minimal” amount of information possible, while at the same time promoting research. Using a real case, an example of an IRB Big Data request submitted to the Johns Hopkins Data Trust Research Subcouncil will be presented. Four panelists from the informatics community will provide 5–10 min comments of their perspectives. A fifth panelist, the assistant dean responsible for human subject protection at Johns Hopkins, will respond. We will then have a brief response by the panelists and an open discussion by the audience.

Introduction

Secondary research using electronic health record (EHR) promises to bring new insights and to take advantage of the massive scale that millions of health records can bring to bear on biomedical- and effectiveness-research questions.¹

Research studies characterized as “data-only” protocols using EHR data are often subject to the Common Rule.² However, IRBs have increasingly focused on the HIPAA privacy rule to inform the reviews of these projects.

A use or disclosure of this information that occurs as a result of, or as “incident to,” an otherwise permitted use or disclosure is permitted as long as the covered entity has adopted reasonable safeguards as required by the Privacy Rule, and the information being shared was limited to the “minimum necessary,” as required by the Privacy Rule.³ [emphasis added]

The current paradigm of IRB oversight of HIPAA-waivered, data-only, studies was born and matured in the context of epidemiologic-style research, where investigators could articulate up front their hypotheses and the data needed to test that hypothesis. The principle of releasing the “minimum” data required fits well with that context.

However, “Big Data”-based research deliberately casts its net more widely in a hypothesis-free exploration of the data specifically because this research is searching for unexpected relationships among variables. Although modern epidemiology needs as much data as it can get it hands on, e.g., to perform high-dimensional propensity score matching,⁴ there are many challenges to the use of clinical data for other purposes.⁵ Machine-learning research, similarly needs tens or hundreds of variables, to discern the features that underlie the data that separate the data in meaningful, if not surprising, ways.⁶ Other branches of clinical research informatics likewise demand access to very broad EHR data in an effort to establish implicit patterns and relationships that are obscured by their complexity, such as discovering clinical pathways.⁷
The potential conflict between the current IRB oversight paradigm and the needs of Big Data researchers is most exercised when the researchers request a large subset of EHR data, including full-text notes, to be housed outside the direct purview of the institutional honest broker, and encourage students or other researchers to explore the dataset. Such a dataset is “pluripotent” in the sense that many research projects can be built off of it, directly opposite to the paradigm of the hypothesis-specific dataset.

The purpose of this panel is to explore the current regime, to assess the challenges derived from the needs of research using pluripotent databases, and to arrive at action items for the informatics community to pursue to overcome the challenges.

Audience members who would most benefit from this panel are researchers aiming to perform Big Data EHR-based research, administrators with oversight responsibility for such projects, and informatics researchers who may be able to address some concerns through technical means.

The Panel Method

Using a case presentation, an example of an IRB Big Data request, based on a proposal submitted to the Johns Hopkins Data Trust Research Subcouncil, will be presented. Four panelists from the informatics community will provide 5–10 min comments of their perspectives. A fifth panelist, the assistant dean responsible for human subjects protection at Johns Hopkins, will respond. We will then have a brief response by the panelists and an open discussion by the audience.

The Case

The researcher—from the Department of Computer Science in the Engineering School, on a separate campus from the School of Medicine—requests “all the data,” including all full-text notes, on 9,000 or so patients admitted to the Johns Hopkins Hospital over 3 years (specific dates provided) for a specific set of operative procedures. The researcher plans to deidentify the data, including date obfuscation and anonymization of the notes. The deidentification will be performed by a vendor with experience in this process and with the Johns Hopkins EHR. The data will be stored on a server managed by the Johns Hopkins cloud-service group, considered “behind the firewall” and used by all research databases employing protected health information. Access to the database will be via a virtual machine (the “SAFE desktop”). While there are 6 general hypotheses (i.e., 6 classes of adverse effects), students will be invited to propose and perform research studies, under guidance of the researcher and other vetted computer science faculty.

The Panel

Jihad Obeid, MD, Division Leader, Biomedical Informatics, Health Systems and Policy, Department of Public Health Sciences, Co-Director, Biomedical Informatics Center, Medical University of South Carolina. Having established the process of EHR data warehouse use for research and garnered its approval by the IRB, he will compare his experience and what has succeeded with the case presented.

William Hersh, MD FACMI, FACP, Professor and Chair, Department of Medical Informatics & Clinical Epidemiology (DMICE), Oregon Health & Science University. An informatics researcher in information retrieval and natural language processing, he will discuss the use of a large corpus of identifiable patient data in a highly secure manner and some additional ideas for how to carry out such research through approaches such as Evaluation as a Service where researchers send systems to the data and are returned aggregate results.

Umberto Tachinardi, MD, MSc, FACMI, Professor, Department of Biostatistics and Medical Informatics, School of Medicine and Public Health, University of Wisconsin, Madison; Associate Dean for Biomedical Informatics. He will provide the Chief Research Information Officer’s perspective.

Tony Solomonides, PhD, Director of Research and Quality Improvement Education, Department of Family Medicine, NorthShore University HealthSystem. As chair of the AMIA Working Group on Ethics, Legal, and Social Issues, he will provide the ethics perspective.
Our Respondent

Megan Kasimatis Singleton, JD, MBE, CIP, Assistant Dean, Human Research Protections and Director of the Human Research Protections Program, Office of Human Subjects Research, Johns Hopkins University School of Medicine. As the institutional representative charged with the oversight of Johns Hopkins Medicine’s IRBs, she will provide the regulatory perspective and response to the issues raised by the panelists.

Discussion Questions

Potential issues to be discussed and for which audience input is desired:

1. What level of granularity of the hypothesis and of specification of the variables for analysis serve as the compromise between the Big Data researcher and the regulator?
2. When is a database “inside” the scope of the honest broker and when not, even in the same institution?
3. What is the minimum level of security required for Big Data research, ranging from physical security to controlled access to audit logs?
4. Who is responsible for vetting and supervising students, when the computing faculty may be unfamiliar with the details of clinical-data governance?
5. Can full-text notes ever be deidentified enough for release to non-clinical students, even under an acceptable security regime?
6. Can other technological approaches, such as Evaluation as a Service, allow research to progress while preserving security?
7. Does “deidentification” even of structured data diminish the power of machine-learning algorithms?
8. If larger, more “open” databases are approved for Big Data research, what alternative oversight mechanisms might be implemented to ensure risks are minimized?

Conclusion

While the regulatory environment is still in flux as the modifications to the Common Rule remain in “administrative review,” action items raised by the panel may provide AMIA in general—and audience members in particular—concrete ways of addressing these pressing issues.

Acknowledgements

Thanks to Ilya Shpitser for raising this concern to the Johns Hopkins Data Trust Research Subcouncil and providing the case. Thanks to Kate Fultz Hollis for helping to assemble this panel.

References

Panel Proposal: Deep Learning for Healthcare - Hype or the real thing?

Jimeng Sun, PhD, Georgia Institute of Technology, Atlanta, GA, USA (moderator)
M. Brandon Westover, MD, PhD, Mass General Hospital & Harvard Medical School
Hong Yu, MS, PhD, FACMI, UMass Medical School, Worcester, MA & Bedford VAMC, MA
Marzyeh Ghassemi, PhD, University of Toronto & Google Verily
David Sontag, Assistant Professor of Electrical Engineering and Computer Science at MIT

Abstract
Deep learning methods refer to variants of neural network models where two or more layers of nodes are integrated with each other. Use of these models is in a renaissance given the substantial advances in the accuracy of neural network models for computer vision, natural language processing and speech recognition. Availability of large scale datasets, advances in high-performance computational devices such as Graphics Processing Units (GPU), and developments of new optimization techniques have permitted efficient learning of deep neural networks for a diversity of application domains.

Deep learning methods are being applied in the healthcare domain:
Automated feature learning: Autoencoders have been used for phenotyping purposes using time-series data. Skip-gram methods were used for learning the representation vectors for medical codes such as diagnosis codes and medication codes. Recurrent networks (RNN) have been used to learn representations from clinical notes. Convolutional networks (ConvNet) also showed great potential for learning effective representations of continuous vital signs.
Accurate predictive models: Training a high-capacity model for accurate prediction is one of the most sought-after direction from many researchers as well as practitioners. ConvNets have shown state-of-the-art performance for detecting retinopathy from retinal images and detecting skin cancer from skin photographs. RNN has also shown great potential for multi-label diagnoses prediction and early detection of heart failure onset.
Synthetic data generation: Generating synthetic electronic health records is another popular direction due to the sensitive nature of private records. Generative adversarial networks is actively being used to generate synthetic claims data and lab measures, or even to guarantee some level of privacy.

However, deep learning models still have many limitations in support healthcare applications:
Interpretation: The deep learning models are often complex black boxes, which are difficult to understand. However, clinical practitioners often prefer and trust simpler and interpretable models. We will discuss how to bridge the gap between model complexity and interpretation.
Causal learning: Great predictive models do not necessarily lead to causal discovery. For clinical applications such as treatment selection, we need to know what treatment change will lead to the best outcome, which often requires identifying causal relations.

Intended Audience: Medical informatic researchers and practitioners.

Topics for discussion:

Deep learning on structured electronic health records (EHR) (by Prof. Sun): Jimeng Sun, PhD, will present new development in interpretable deep learning models for health outcome prediction using structured EHR data. Dr. Sun will present multiple deep learning applications such as heart failure prediction with Sutter Health and patient complexity prediction with Children’s Healthcare of Atlanta. In particular, he present the following 1) accuracy prediction via temporal modeling, 2) representation learning through multi-layer embedding, 3) interpretable deep learning model via temporal attention model.

Deep learning for clinical neurophysiology: epilepsy, neuro-critical care, and sleep. M. Brandon Westover, MD, PhD, will discuss using deep learning in large electroencephalography (EEG) datasets to
advance care for patients with neurophysiological problems. He will describe: development of the world’s largest set of training data for epileptic spike detection, and for seizures in ICU patients; and creation of a sleep-staging algorithm that performs as well as human experts on real-world polysomnography recordings, and how this is making it possible to characterize a patient’s “brain age”.

**Deep learning for clinical NLP:** Prof. Hong Yu will describe innovative recurrent neural network models (e.g., RNN and memory-augmented neural network models) we developed and their applications to electronic health record notes clinical event detection and relation identification. We will also describe the learning behavior of neural network models, supporting that their learning behavior is similar to the behavior of human.

**The limits of deep learning in healthcare:** David Sontag, Ph.D., will give a broad overview of the use of deep learning in healthcare with the goal of emphasizing examples where deep learning has significant potential, and conversely characterizing where deep learning *does not* provide significant improvement over simpler machine learning algorithms. Examples will range from machine learning in radiology and pathology to predictive modeling on health insurance claims and physiological time-series.

**Phenomics is the new genomics:** Marzeyeh Ghassemi, PhD. will present an overview of the ways that in-patient ICU records spanning text, signals and demographics can be used predict important clinical events, and understand the need for treatments. She will discuss the improvements and drawbacks that deep learning provides to learning in clinical settings with heterogeneous healthcare records.

**Timing for the panel**

Deep learning has shown great successes in many domains with big data sets. However, its impact on healthcare applications is still very limited. As diverse large biomedical data are being collected, people can foresee potential success of deep learning in healthcare. We assemble a group of pioneers in researching advanced deep learning models for healthcare problems to provide their perspectives in this hot topic and to try to answer whether deep learning is a hype or a real thing for addressing healthcare problems.

**Discussion questions**

- What are the major successes of deep learning in healthcare?
- What are the challenges in using deep learning models for diverse biomedical data sources?
- How to provide interpretable models in the era of deep learning?
- What are the common mistakes in using deep learning for addressing healthcare problems?

**Panel lineup:**

**Jimeng Sun (moderator),** Ph.D., is an Associate Professor of School of Computational Science and Engineering and a co-director of center of health analytics and informatics at College of Computing at Georgia Institute of Technology. His research focuses on health analytics and medical informatics, especially in applying large-scale predictive modeling and computational phenotyping on biomedical applications. He has published over 100 papers, filed over 20 patents (5 granted). He has received SDM/IBM Early Career Data Mining Research Award 2017, ICDM Best Research paper in 2008, SDM Best Research Paper in 2007, and KDD Dissertation Runner-up Award in 2008. Dr. Sun received his B.S. and M.Phil. in Computer Science from Hong Kong University of Science and Technology in 2002 and 2003, and PhD in Computer Science in Carnegie Mellon University in 2007. Prior to joining Georgia Tech, He was a research staff member at IBM TJ Watson Research Center.

**M. Brandon Westover,** MD, PhD, completed medical training and a PhD degree in physics at Washington University School of Medicine in St. Louis. He is an Assistant Professor of Neurology at
Harvard Medical School and a neurologist specializing in epilepsy and clinical neurophysiology at the Massachusetts General Hospital (MGH) where he directs the MGH Critical Care EEG Monitoring Service. Clinically, he runs a service that provides electroencephalography (EEG) monitoring to help care for patients with acute neurological conditions such as delirium, anoxic brain injury, status epilepticus, and delayed cerebral ischemia following subarachnoid hemorrhage. His research interests include automated methods for interpreting clinical EEG data, closed-loop control of sedation and analgesia, biomedical informatics, medical decision making under uncertainty, and the neurophysiology of pain, sedation, and delirium in critically ill patients.

**Hong Yu** Ph.D., Professor, Department of Quantitative Health Sciences, UMASS Medical School Professor Yu's research interests are in the areas of information retrieval, extraction, natural language processing, summarization, human-computer interaction, with a focus on biomedical applications. She has led the development of many systems, including the biomedical question answering system AskHERMES (http://www.askhermes.org), the adverse drug event pharmacovigilance system (http://www.ADErepository.org), the innovative biomedical figure search system (http://figuresearch.askhermes.org), and the NoteAid systems that help patients comprehend electronic medical records (http://ClinicalNotesAid.org).

**Marzyeh Ghassemi**, Ph.D., Visiting Researcher with Google's Verily and a post-doc with Dr. Peter Szolovits in MEDG; in Fall 2018 she will be joining the University of Toronto as an Assistant Professor in Computer Science and Medicine, affiliated with the Vector Institute. She has clinical collaborations at Beth Israel Deaconess Medical Center and Massachusetts General Hospital. She has organized the NIPS 2014 Women in Machine Learning Workshop, the NIPS 2016 Workshop on Machine Learning for Health, and MIT's first Hacking Discrimination event. Her research interests are centered around creating and applying machine learning algorithms towards improved prediction and stratification of relevant human risks.

**David Sontag**, Ph.D., Assistant Professor in the Department of Electrical Engineering and Computer Science (EECS) and Hermann L. F. von Helmholtz Career Development Professor in the Institute for Medical Engineering and Science (IMES). He is also a principal investigator in the Computer Science and Artificial Intelligence Laboratory (CSAIL). Sontag's research focuses on machine learning and artificial intelligence; at IMES, he leads a research group that aims to use machine learning to transform health care. Previously, he was an assistant professor in computer science and data science at New York University's Courant Institute of Mathematical Sciences and a postdoctoral researcher at Microsoft Research New England. Dr. Sontag received the Sprowls award for outstanding doctoral thesis in Computer Science at MIT in 2010, best paper awards at the conferences Empirical Methods in Natural Language Processing (EMNLP), Uncertainty in Artificial Intelligence (UAI), and Neural Information Processing Systems (NIPS), faculty awards from Google, Facebook, and Adobe, and a NSF CAREER Award. Dr. Sontag received a B.A. from the University of California, Berkeley.

We affirm that all panel participants have agreed to participate and have contributed to the preparation of this document.
Frontiers of Clinical Information Extraction: Current Progress in Medication and Adverse Drug Event Detection from Electronic Health Records

Hong Yu, Abhyuday Jagannatha, Feifan Liu, Weisong Liu

Abstract:

Recent deep learning has shown superior performance in many artificial intelligence applications, including clinical information extraction. However, developing innovative high-performance information extraction approaches using electronic health record narratives remains a challenging natural language processing (NLP) task. This workshop aims to bring together biomedical researchers, clinicians and industry professionals to discuss methodological advances in clinical information extraction. Specifically, over 10 participant teams of NLP Challenges for Detecting Medication and Adverse Drug Events (MADE1.0) will present their novel design, methodologies as well as state-of-the-art results using over 1,000 expert-annotated EHR notes as a common ground truth gold standard.

Intended Audience: Medical informatics and NLP researchers and practitioners

Number of Anticipated Attendees: 80

Topics for Discussion and Timing:

A. Advanced methods in clinical NLP and their applications in clinical knowledge discovery: Overview of current progress in clinical information extraction and the MADE1.0 results (30 mins)

B. Presentations of selected MADE1.0 participant teams (60 mins)
Improving Interoperability between Registries and EHRs

Seth Blumenthal, MBA; PCPI, Chicago

Abstract

National performance measurement needs clinical data that track the performance of multidisciplinary teams across episodes of care. Clinical registries are ideal platforms for this work due to their capture of structured, specific data across specialties. Because registries collect data at a national level, and registry data are captured in a consistent structure and format within each registry, registry data are useful for measurement and analysis “out of the box”. Registry business models are hampered by the cost of collecting data from EHRs and other source systems and abstracting or mapping them to fit registry data models. The National Quality Registry Network (NQRN) has launched Registries on FHIR, an initiative to lower barriers to achieving semantic interoperability between registries and source data systems. In 2017 Registries on FHIR conducted an information gathering campaign to learn where registries want better interoperability, and how to go about improving it.

Introduction

PCPI is a Chicago-based, national nonprofit organization focused on performance improvement in the health care delivery system to improve health outcomes. (1) The National Quality Registry Network (NQRN) is a PCPI program that supports clinical registries and promotes their use for performance improvement, measurement and other uses. (2) Clinical registries are organized systems that collect data on patients from pre-determined populations for specific purposes. (3) Registries are important platforms for performance measurement that supports value-based payment programs. (4) A Qualified Clinical Data Registry (QCDR) is an entity that collects clinical data from clinicians participating in the CMS Quality Payment Program, and submits the data to CMS on their behalf for the purpose of program participation. (5) Registries collect data from a number of sources and through a combination of direct, manual and automated input methods. Aside from direct manual clinician input, typically through a secure web interface, common registry source data systems include EHRs, clinical information systems, medical devices and patient portals. The data that registries need to capture from these sources vary in form, structure and meaning, requiring mapping to registry data models, and/or manual clinical chart abstraction into registries. This mapping and chart abstraction work adds to the cost of operating and participating in registries. (6)

In 2016 the registry community through the NQRN program asked for help in improving registry interoperability, primarily to reduce these costs and related burdens, but also to lower barriers to implementing performance measures that span multiple clinical specialties and thus use data from multiple registries. In response, PCPI has launched Registries on FHIR, a registry interoperability initiative and a coalition of society registry stewards, registry and EHR vendors, informaticists and others. The initiative began its work with a fact-finding campaign which resulted in a list of priority areas for potential registry interoperability projects. The campaign also informed the development of an interoperability solution process that begins with a clinical scenario or workflow and results in the generation of technical documentation that supports implementing the interoperability solution in registries and source data systems. The purpose of this paper is to inform the informatics community about what we found. Due to the high cost of obtaining the needed data and the increasing use of registries as measurement platforms that inform value-based payment programs, there is an urgent need to reduce registry data acquisition costs. Improving the ability of registries and source data systems to semantically interoperate is a high priority for these societies and for health care.

Methods

The Registries on FHIR information gathering campaign was designed to increase understanding of registry interoperability needs and understand recommended approaches for developing and testing solutions to interoperability problems. A survey and a set of interview questions were developed for identifying potential interoperability projects. The questions intended for registries and registry vendors asked about problems faced that could potentially be reduced or eliminated with better registry interoperability, the motivations for wanting those problems solved, relevant clinical or operational scenarios, data needs, and barriers that needed to be overcome in
order to solve. The interoperability solution process interviews were conducted informally, without working from a set of specific questions.

The registry interoperability needs survey was conducted in June 2017 using SurveyMonkey. The survey was sent to all 2017 QCDRs where a contact email address was found; emails were located for 50 QCDRs. (5) No follow up activities were conducted after the initial survey email was sent. Interviews were conducted in May and June with the registries and registry vendors participating in Registries on FHIR, which were a convenience sample. Four national clinical registries and six registry vendors were interviewed by phone.

Interviews were then conducted with informaticists, including those affiliated with standards-development organizations, industry trade groups and independent consultants in data standards and interoperability. In these interviews, we described the registry interoperability challenge in general and asked for advice on how to solve interoperability problems from a registry perspective.

**Results**

16% (8/50) of the QCDRs responded to the survey, and an additional 4 were interviewed, for a total of 24% of the 2018 QCDRs. 6 registry vendors were interviewed. Registries reported the greatest level of interest in solving problems related to their specific workflows, and a secondary interest in general interoperability problems. The areas identified are listed in Table 1.

**Table 1: Registry Interoperability Priority Areas**

<table>
<thead>
<tr>
<th>Area</th>
<th>Reporter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Develop common data element standards for data that are used in many registries e.g., demographics, vitals, pain, Activities of Daily Living (ADL) score</td>
<td>Registries and registry vendors</td>
</tr>
<tr>
<td>Develop standardized information models that support needs common across many registries e.g., QCDR measures, measure sets needing harmonization, cross-cutting measures</td>
<td>Registries and registry vendors</td>
</tr>
<tr>
<td>Develop nationally-accepted specialty-specific data standards e.g., for pathology specimen collection, and make them generally available</td>
<td>Registries</td>
</tr>
<tr>
<td>Develop Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR) implementation guides that incorporate existing data standards i.e., American College of Cardiology data standards</td>
<td>Registries</td>
</tr>
<tr>
<td>Improve patient matching; particularly useful in specialties that have a lot of overlap i.e., anesthesia</td>
<td>Registry vendors</td>
</tr>
<tr>
<td>Identify specific instances of concepts in registries e.g., a system to standardize the way anything is tagged or identified in registries</td>
<td>Registry vendors</td>
</tr>
<tr>
<td>Develop technical interoperability standards or guidance for registries</td>
<td>Registry vendors</td>
</tr>
</tbody>
</table>

In interviews with informaticists, a number of steps were consistently identified. The first step is to identify the clinical or operational scenario that may benefit from improved interoperability. This could be expressed in the form of a case or user story, a clinical pathway, or other similar descriptive format. Regardless of the format used, in this step the information needs and flows should be identified and documented. Next an information model should be identified or created that documents and abstracts the data and information needs at a level that allows the needed information to be stored and communicated. Information models capture the semantics of the scenario, and are technology-agnostic. (7) In order for information models to be created that are generalizable, the data elements or concepts they represent need to be agreed upon. As an example, an information model involving smoking cessation
will use or define the needed concepts in a consistent manner, so that the same basic data elements are used in a way that preserves meaning across implementations. Depending on the relevant clinical or technology domains involved in the scenario, an HL7 Domain Analysis Model (DAM) may be available or need to be created or updated. DAMs provide a detailed definition of a subject of interest, such as a particular clinical process, at a level of detail useful to domain experts and non-technical stakeholders who have an interest in seeing that domain’s semantics expressed in a standardized way in technology implementations. (8) Additionally, HL7 is collaborating with the Healthcare Services Platform Consortium (HSPC) and other organizations in launching the Clinical Information Interoperability Council (CIIC). The CIIC intends to create an information model repository and setup a governance and support structure for the development and maintenance of clinical information models. (9)

Once the needed information models are identified or created, then technology-specific implementation guidance can be created i.e., in the form of a FHIR Implementation Guide or an Integrating the Healthcare Enterprise (IHE) profile. (10) (11) With this guidance available at least in draft form, an interoperability demonstration can be conducted. In this case, with a registry interoperability problem under investigation, the implementation guide would be used to update the registry, source data systems and possibly also other systems and processes used to support information exchange. The performance of the systems in question may then be measured before and after the change to hopefully demonstrate that the improved performance justifies the work performed. The demonstration work will also provide feedback that can be used to enhance the draft implementation guides as they are made available for general use in their respective standards-development organizations’ standards libraries.

Discussion

Interoperability, even that focused on registries covers a wide area and it is not realistic to expect that a single methodology or set of specific steps can be used to solve registry interoperability problems in general. However, a number of themes came out of the interviews, from which a broad outline of a process could be constructed. Fundamentally, any interoperability in health care is important because of the use of health information technology in most clinical or administrative workflows.

Value-based payment models are driving an increasing need for measurement and comparative analytics that cross multiple clinical specialty or organization boundaries. The current lack of interoperability of health care data and content impairs the industry’s ability to realize better value (decreased costs and increased benefits) for populations and individual patients. To achieve better value, we need to realize the following experiences related to interoperability. Anyone taking care of a patient, including the patient, must have appropriate access to all pertinent data from all sources to make the best-informed decisions. Technical interoperability allows access to data. To deal with the complexity, volume, and velocity of health data, patients, providers, and population health practitioners need decision support from health information technology to understand and intervene in patient care and population health. Semantic interoperability allows health IT to reliably understand the meaning of the information contained in the data, and thus to reliably use them in decision-support algorithms. (12)

In addition to access to data and understanding of the information contained within the data, we need health information technology to capture and facilitate the pertinent health care processes. In order to manage health care services at the individual and organizational levels, an understanding is needed of which processes are effective or wasteful. Process interoperability exists when information systems can make useful recommendations as to the processes to recommend for patients, and thus assist healthcare providers in reliably understanding and implementing a collaborative plan of care.

The health care industry has so far mostly focused on technical interoperability, also known as syntactic or structural interoperability. Health information exchanges and standards such as FHIR are enabling systems to exchange data to support data access. Adoption of standards for technical interoperability is increasing, but the infrastructure that supports data liquidity lags.

Standards that allow meaningful semantic and process interoperability either do not yet exist or are not yet feasible to apply and thus have not been widely adopted in health IT. Almost every provider organization has a different (and largely incompatible) approach to semantic interoperability. Almost no healthcare institution uses standards to represent processes. The lack of interoperability on these levels is a serious problem, as other industries derive most
of their value from managing standardized decisions and processes. Robust semantic and process interoperability is critical in medicine, as cooperative care is increasingly the norm, and no one organization can create content for decision and process support that accounts for the best clinical evidence. Semantic and process interoperability is required for robust markets for content and cooperative care that improves health outcomes.

Clinical Registries
Clinical registries collect clinical and administrative data on patients and the care they receive. The primary focus of registry data collection is capturing information about clinical treatment and related patient health outcomes. The resulting registry datasets then support treatment, payment, quality improvement, benchmarking and research, as well as other uses.

Registries specialize in collecting specific data elements of importance to the clinical domains in which they focus. However, certain types of data e.g., demographics are likely collected in all or most registries. Other data such as cardiology clinical data may be collected not only in cardiology-specific registries, but also in other registries capturing data on patients undergoing treatment in other clinical domains for which cardiovascular disease is a factor e.g., cardiology data elements. These may exist in registries focused on surgery, endocrinology and other clinical domains, and despite the commonality of data, most registries are not using a common format for these data elements. There is currently no standard that defines how common data elements should be defined and captured in registries. Some registry stewards have published standards for their specialty-specific data elements, but those standards have not been widely adopted.

EHRs are a common data source for registries. Individual EHR implementations are typically contained within a single health system or provider organization. There are over 1,000 health systems in the United States, and health systems use a number of different vendors for their EHRs. EHR data vary in content, form and structure across EHR implementations, even those from the same vendor. In response, some institutions have implemented data warehouses that provide a single place in which to store data from the multiple data systems that are independently capturing, measuring clinician performance and reporting to evaluating organizations on behalf of the institution. But in looking at data nationally, even health system data warehouses are still very numerous and their data are also not standardized from one data warehouse to another.

Although there is tremendous value contained within individual registry data, the cost of collecting the data and capturing them in registries in a way that preserves the meaning of the information contained within the data i.e., semantic interoperability, continues to be high. The current work required to extract data from EHRs is one-to-one, duplicative and not scalable. This drives up cost, which is high even when looking at data across multiple clinical domains, and thus across multiple registries is even feasible.

Registries capture data in highly structured formats, with a high level of standardization between different patients, case reports and participating clinicians within a single registry, but there is a broad lack of standardization, at all levels of conceptualization, across registries. Registries do not report uniform data (in structure or semantics) so anyone taking a broad look at health outcomes e.g., health systems using analytics platforms, and CMS, must harmonize and normalize the registry information they use at a semantic level to accurately and fairly measure and report on the results of clinical processes and patient health outcomes.

The work required to achieve data liquidity across registries, EHRs and other systems is considerable. Due to the lack of structure and standardization of EHR data, a significant percentage of the data required to calculate performance i.e., performance measures for CMS reporting programs, must be entered in registries via manual clinical chart abstraction. In order to extract data automatically from EHRs, warehouses and other source data systems into registries, custom interfaces between the registry and each individual source data system implementation must be created. These interfaces map the source data into the format and structure required to perform the measure calculations in a way that clinician comparisons on a national level can be made with accuracy and validity. These interfaces add cost and must be implemented for each organization that participates in the registry.
Conclusion
With the continued emphasis in value-based payment, and on measuring health outcomes across organizational boundaries over varying periods of time, demand for information from registries is increasing. But due to the high cost, it is often not feasible to capture all of the data registries need. Additionally, despite the benefits for national performance measurement provided by registries, it is difficult to aggregate data across multiple registries to come up with uniform data that are valid for measuring and improving performance and health outcomes across a wide variety of patient populations, diseases and conditions and the clinical specialties that provide care for those patients.

These and other factors have created a burning platform for increased semantic interoperability between registries and the health IT they exchange data with, so that our return on investment in data collection, measurement and improvement is maximized.

Registries on FHIR is an initiative that aims to improve registry interoperability, in particular interoperability between registries and source data systems such as EHRs. The initiative will begin by examining the current state, and then identifying registry interoperability problems to solve. The scope from which initial projects will be identified includes technical and semantic interoperability between registries and EHRs, and may include work on any relevant level of abstraction from data transport standards down to individual data element standardization. In later phases, the initiative may expand to include other areas such as linking between registries and utilizing semantic web platforms to standardize the way registries, EHRs and clinical applications interoperate.

For each project that Registries on FHIR takes on, the goal will be to demonstrate increased data liquidity within a test group; to show that a state of lower effort or cost to operate registries or capture high-priority data has been achieved by using the infrastructure that has been developed in this project. If applicable the project will then make the infrastructure available for use through NQRN guidance and/or nationally- or internationally-accepted bodies such as standards development organizations.

We will promote any success achieved, and then attempt to use the output each project as a catalyst for further work to affect broader adoption of these and other data standards among clinical registries and other clinical information systems that serve as data sources for or users of registry data.

References


http://www.hl7.org/fhir/?ref=learnmore.

https://www.ihe.net/Profiles/.


What’s in a Note? Unpacking Predictive Value in Clinical Note Representations

Willie Boag, B.S. 1*, Dustin Doss, S.B. 1*, Tristan Naumann, M.S. 1*, Peter Szolovits, Ph.D. 1
1 Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract

Electronic Health Records (EHRs) have seen a rapid increase in adoption during the last decade. The narrative prose contained in clinical notes is unstructured and unlocking its full potential has proved challenging. Many studies incorporating clinical notes have applied simple information extraction models to build representations that enhance a downstream clinical prediction task, such as mortality or readmission. Improved predictive performance suggests a “good” representation. However, these extrinsic evaluations are blind to most of the insight contained in the notes. In order to better understand the power of expressive clinical prose, we investigate both intrinsic and extrinsic methods for understanding several common note representations. To ensure replicability and to support the clinical modeling community, we run all experiments on publicly-available data and provide our code.

Introduction

Electronic Health Records (EHRs) contain an abundance of data about patient physiology, interventions and treatments, and diagnoses. The amount of data can be overwhelming in the intensive care unit (ICU), where patients are severely ill and monitored closely. In this setting, it can be difficult to reconcile data from multiple sources; instead, care staff rely on clinical notes to provide short summaries that capture important events and results. This unstructured, free text thus contains important observations about patient state and interventions, in addition to providing insight from caregivers about patient trajectory.

The secondary use of EHR data in retrospective analyses facilitates a better understanding of factors, such as those contained in clinical notes, that are highly predictive of patient outcomes. 1,2,3,4 Additionally, the free-text nature of clinical notes means that data extraction does not rely heavily on each EHR’s implementation, making methods for clinical notes portable across different EHRs. However, there are many ways to represent the information contained in text, and it is unclear how to best represent clinical narratives for the purpose of predicting outcomes.

Many efforts to leverage clinical notes for outcome prediction focus on improving the performance of a final prediction task. 1,3,5,6 Post-hoc feature analysis can assist in discovering those features that are most predictive, but it provides only a partial solution toward improving our understanding. We would like to know what facts and derived features matter most in affecting the predictive abilities of the models we build from them. This will allow us not only to improve performance but to understand what representations of the identified features are most useful.

For example, although a patient’s EHR-coded race and social history may help to identify a Gonorrhea infection accurately, 7 if we are trying to use text analysis tools to make such an identification, we would like to know if those tools are able to determine a patient’s race and social history accurately from the notes. While it is seemingly counter-intuitive to predict EHR-coded information using clinical notes, doing so provides insight into what is, and isn’t, reflected in a given note’s representation. Such awareness is important when designing representations for downstream prediction tasks because it exposes assumptions about what sophisticated models may be able to accomplish.

Toward the goal of understanding and improving note representations for downstream prediction performance, we consider several common representations and evaluate them on a variety of tasks. We explore performance on “easy” tasks, such as age, gender, ethnicity, and admission type, each of which are readily accessible as EHR-coded data. Additionally, we use the same models on common prediction tasks, such as in-hospital mortality and length of stay. We show that 1) no single representation outperforms all others, 2) a simple representation tends to outperform more complex representations on “easy” tasks while the opposite is true for “common” prediction tasks, and 3) some seemingly “easy” tasks, such as ethnicity, are difficult for all of the representations considered.

* Authors contributed equally.
Figure 1. An example clinical note. The age, gender, and admitting diagnosis have been highlighted. Also note, that descriptions such as “status worsening” suggest deterioration and possible in-hospital mortality.

Related Work

Work leveraging clinical notes for prediction can be broadly categorized into those focusing on clinical prediction tasks and those focusing on the representation of text.

Clinical Prediction Tasks: Several existing works have demonstrated the utility of clinical narratives in forecasting outcomes. A standard approach for converting narrative prose to structured vector-based features has used unsupervised topic modeling to represent each note as a distribution over various topics. Lehman et al.⁶ and Ghassemi et al.¹⁻⁵ use note-derived features in a framework to predict mortality. In recent work, Caballero Barajas and Akella² use generalized linear dynamic models on top of latent topics to detect an increase in the probability of mortality before it occurs. Luo and Rumshisky⁸ use a supervised topic modeling approach to improve prediction of 30-day mortality. Grnarova et al.⁴ use convolutional neural networks (CNNs) to construct document representations for the task of mortality prediction. Although the authors do not perform this prediction task in a time-varying manner, their results show that both doc2vec⁹ and their CNN approach improve performance relative to a topic representation. Further, Cohen et al.¹⁰ explore the use of redundancy-aware topic models in order to combat the prevalent issue of copying notes forward in a patient’s clinical record; however, they do not apply this model in a downstream prediction task. Similarly, Pivovarov et al.¹¹ explore the use of topic models in the discovery of probabilistic phenotypes, but do not use these phenotypes toward predictions.

Text Representations: In the general domain, it has been observed that simple models with access to much data often outperform complex models with access to less data. Banko and Brill¹² observe this effect directly in the general natural language processing domain, noting that many methods continue to be optimized on small datasets and prove ineffective when applied to datasets orders of magnitude larger. Similarly, Halevy et al.¹³ discover that “for many tasks, words and word combinations provide all the representational machinery we need to learn from text.” Previously, limited access to clinical narratives have made this observation less applicable to the clinical domain.

Data

This work uses data from the publicly-available Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) database, version 1.4.¹⁻¹⁴ MIMIC-III v1.4 contains de-identified EHR data from over 58,000 hospital admissions for nearly 38,600 adult patients. The data were collected from Beth Israel Deaconess Medical Center from 2001–2012. A typical clinical note might look like the one shown in Figure 1, which shows the radiology report of a 64-year-old patient with poor respiratory status.
A patient’s time in the ICU generates a sequence of timestamped notes. Each of the methods described transforms the sequence of notes into a fixed-length vector representing the ICU stay.

We consider only adult patients, older than 15 years old, who were in the ICU for at least 12 hours. Young patients are excluded as they typically exhibit different physiology from an adult population. Further, we include only each patient’s first ICU stay, thus precluding training and testing on data from the same patient. Due to recording and measurement issues in the database, we exclude any ICU stays that do not conform to the common sense ordering of

\[ \text{hosp\_admission} \leq \text{icu\_intime} \leq \text{icu\_outtime} \leq \text{hosp\_dischtime} \]

Finally, we consider Nursing and Nursing/Other, Radiology, and Physician notes, because other categories occurred relatively infrequently. For each ICU stay, we extract the first 24 notes (or fewer if the stay has fewer notes). These criteria result in 29,979 unique ICU stays, an equivalent number of patients, and 320,855 notes. The dataset is randomly divided into a 7:2 train/test split.

As “easy” prediction tasks, we extract several coded variables for each patient that remain constant throughout the stay, including: age, gender, ethnicity, and admission type. In addition, we also retrieve “common” clinical outcomes and findings during the stay, such as: diagnosis, length of stay, and in-hospital mortality. We then try to predict these characteristics and outcomes from different representations of the text notes.

As observed in replication studies, one of the central obstacles in replicability — even for work done on public datasets — is that descriptions of data cleaning and preprocessing are often inadvertently underspecified. Therefore, we make our code publicly available.

Methods

MIMIC-III v1.4 contains de-identified clinical notes. In preprocessing these notes, tags indicating de-identified protected health information are removed. Phrases written entirely in capital characters are then replaced by a single token, effectively coalescing common structural elements; for example, the section heading “RADIOLOGIC STUDIES” would be replaced with a single token. Additionally, regular expressions for common age patterns are used to replace all specified ages with symbols binned by decade to ensure relevant age information is not lost. Finally, we remove all non-alphanumeric tokens, and normalize all remaining numbers to a single number token.

For each word, we compute the number of unique patients who have a note containing that word — this is the “document” frequency. For each note, we compute the term frequency-inverse document frequency, or tf-idf of every word and keep the top-20 words of that document. Thus a patient’s stay is represented as an ordered list of filtered bags-of-words.

The following subsections describe several approaches to aggregate each patient’s multiple note vectors into one fixed-size patient vector that summarizes their stay in the ICU, as illustrated in Figure 2.

\[ ^1 \text{Code available at } \text{http://www.github.com/wboag/wian.} \]
**Bag of Words**

Bag of words (BoW) is one the simplest methods for creating vector representations of documents. Using the top-20 tf-idf words from each note produces a vocabulary of size $|V| = 17,025$ words. In this representation, the *patient vector* is a $|V|$-dimensional sparse, multi-hot vector. If a word appears in any of the notes for a given patient, then the corresponding dimension for that word is “on” in the resulting *patient vector*.

Bag of words presents a strong baseline representation for downstream predictive tasks. In this work, its strength is a result of its high dimensionality relative to other models: by reducing the representations into a smaller, denser space, other models may inadvertently throw out information with predictive value. More specifically, we expect that bag of words will perform well on tasks that involve the prediction of categories which may be directly represented by single words in their notes. For example, we would expect a note which frequently contains the word “male” to correctly identify the patient as male.

**Word Embeddings**

Due to the immense success of word2vec in recent years, we embed words and documents into a dense space in order to accommodate soft similarities. We train clinical word vectors using the publicly-available word2vec tool‡ on 129 million words from 500,000 notes taken from MIMIC-III. Hyperparameters were specified using Levy et al.16 as a reference: 300-dimensional SGNS with 10 negative samples, a min-count of 10, a subsampling rate of 1e-5, and a 10-word window. These clinical embeddings are available for public use on the MIMIC-III Derived Data Repository.§

As shown in Figure 3, we create a note representation by aggregating the top tf-idf words in the document. With these top words, we look up each of their word2vec embeddings (blue) and collapse them into a final vector using elementwise -max, -min, and -average. We apply the same aggregation scheme (max, min, and average) to collapse the patient’s list of document vectors into one fixed-length *patient vector*.

**Recurrent Neural Network**

One problem with the approaches described above is that they all ignore temporal ordering of the documents. That

---

‡ Code available at https://github.com/tmikolov/word2vec.
§ Data available at https://physionet.org/works/MIMICIIIDerivedDataRepository/.

---

**Figure 3.** How the embedding for a single document is built by combining constituent word embeddings.

**Figure 4.** The many-to-one prediction task for the LSTM, in which a document representation is fed in at each timestep, and it makes a prediction (e.g. diagnosis) at the end of the sequence.
is to say, they fail to track the progression of a patient’s state over time during the ICU stay. One solution to this limitation is to use a sequence-based model. We use a Bidirectional LSTM network, which has proven to be very effective at modeling temporal sequences. For a more fair comparison, we build the list of document vectors for each patient in the same way that was done for word embeddings. These document vectors are then fed into the LSTM one document per timestep.

Our LSTM was implemented in Keras using an Bidirectional LSTM with 256 hidden units, a dropout rate of 0.5, and a 128-unit fully connected layer immediately before the output label softmax. Models were trained for 100 epochs.

**Experimental Setup**

The principal aim of this work is to better understand what information is captured by various representations of clinical notes. Because most of the derived representations have non-interpretable dimensions from the embedding process, we cannot look for correlations between individual dimensions and our queries. Instead, we develop a prediction framework to determine whether a particular representation has encoded the necessary information to predict the correct query against alternative values.

We consider the following prediction tasks modeling clinical states and outcomes:

1. **Diagnosis.** We filter down to patients with one of the 5 most common primary diagnoses and predict: *Coronary Artery Disease, Pneumonia, Sepsis, Intracranial Hemorrhage, and Gastrointestinal Bleed.*

2. **In-Hospital Mortality.** Binary classification of whether the patient died during their hospital stay.

3. **Admission Type.** Binary classification of *Urgent* or *Elective.*

4. **Length-of-Stay.** Three-way classification of whether patients stayed in the ICU for *Less than 1.5 days, Between 1.5 and 3.5 days,* and *longer than 3.5 days.*

However, we are also interested in whether the notes are able to capture basic demographic information:

1. **Gender.** Binary classification of *Male* or *Female.*

2. **Ethnicity.** Binary classification of *White* or *Non-White.*

3. **Age.** Three-way classification of age as *Less than 50 years old, Between 50 and 80 years old,* or *Older than 80 years old.*

While these tasks reflect those commonly found in research, we use them to evaluate our representations rather than as clinically-actionable targets. For example, it might be noted that a single patient can suffer from multiple conditions, but here we consider only their primary diagnosis. Similarly, the ranges for age and length-of-stay are reasonable, but would need to be tailored in other conceivable applications. In both cases, however, these choices serve to highlight the types of information each representation is capturing.

Binary classification tasks are evaluated using AUC, while multi-way classification tasks are evaluated using the macro-average F1-score of the different labels. Predictions are made for bag-of-words and word embedding representations using a scikit-learn support vector classifier with linear kernel. Predictions are made for the LSTM using a softmax layer.

**Results**

Performance for the 7 classification tasks using the 3 representation models are shown in Table 1 (binary classifications) and Table 2 (multi-way classifications). In general, our findings match our expectations: while a complex model tends to do well for “downstream” tasks involving reasoning, such as diagnosis and length-of-stay, it struggles to compete with a simpler model in token-matching tasks like age and gender.
Table 1. AUCs for the binary classification tasks.

<table>
<thead>
<tr>
<th></th>
<th>in-hospital mortality</th>
<th>admission type</th>
<th>gender</th>
<th>ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.821</td>
<td>0.883</td>
<td>0.914</td>
<td>0.619</td>
</tr>
<tr>
<td>Embeddings</td>
<td>0.814</td>
<td>0.873</td>
<td>0.836</td>
<td>0.580</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.777</td>
<td>0.870</td>
<td>0.837</td>
<td>0.533</td>
</tr>
</tbody>
</table>

Table 2. Macro-average F1 scores for the multi-way classification tasks.

<table>
<thead>
<tr>
<th></th>
<th>diagnosis</th>
<th>length of stay</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoW</td>
<td>0.828</td>
<td>0.724</td>
<td>0.635</td>
</tr>
<tr>
<td>Embeddings</td>
<td>0.828</td>
<td>0.730</td>
<td>0.544</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.836</td>
<td>0.758</td>
<td>0.450</td>
</tr>
</tbody>
</table>

Specifically, the bag-of-words (BoW) model performs best at predicting so-called ‘common-sense’ tasks: age, gender, and (less significantly) ethnicity, for which there are words which almost directly predict the labels. In contrast, the LSTM model outperforms BoW on tasks more related to clinical reasoning: diagnosis and length of stay, for which we expect the temporal information to be important in predictions. Embeddings serve as a halfway between BoW and LSTM; while the method does not leverage a temporal sequence, this experiment allows us to untie the pre-trained word vectors from the temporal dynamics of the LSTM. In doing so, we see that the embeddings typically perform very competitively against BoW, but the LSTM is able to leverage them further.

Discussion

As shown in Table 1 and Table 2, the different models exhibit varied performance across tasks with no consistent winner. Bag-of-words tends to do well on tasks where a single word, or a few words, are strongly associated with prediction categories. Notably, bag-of-words is much better at predicting age. This is likely because the normalized, per-decade age tokens created during preprocessing are, of course, strongly associated with predicting age. The LSTM, on the other hand, had a difficult time distinguishing between the age token embeddings since all age tokens fall nearby one another within the embedding space, as shown in Figure 5.

For these tasks, bag-of-words provide a strong baseline because some standard demographic information, such as age and gender, are typically specified in the notes. However, it is precisely because of their frequency of occurrence that information retrieval methods, such as tf-idf, underestimate their importance. Recall that tf-idf reduces the score of

![Figure 5. PCA 2-D projection of the word embeddings. Vectors of the special age tokens are colored red. Note that these tokens cluster close together in the embedding.](image-url)
Table 3. Most predictive words for gender

(a) Male

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>1.4012</td>
</tr>
<tr>
<td>he</td>
<td>1.0589</td>
</tr>
<tr>
<td>wife</td>
<td>0.9953</td>
</tr>
<tr>
<td>male</td>
<td>0.7956</td>
</tr>
<tr>
<td>his</td>
<td>0.6772</td>
</tr>
<tr>
<td>prostate</td>
<td>0.2435</td>
</tr>
<tr>
<td>prop</td>
<td>0.1965</td>
</tr>
<tr>
<td>ofm</td>
<td>0.1850</td>
</tr>
<tr>
<td>hematuria</td>
<td>0.1816</td>
</tr>
<tr>
<td>distention</td>
<td>0.1756</td>
</tr>
<tr>
<td>trauma</td>
<td>0.1748</td>
</tr>
</tbody>
</table>

(b) Female

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>she</td>
<td>1.0176</td>
</tr>
<tr>
<td>woman</td>
<td>0.9051</td>
</tr>
<tr>
<td>her</td>
<td>0.7561</td>
</tr>
<tr>
<td>husband</td>
<td>0.7004</td>
</tr>
<tr>
<td>breast</td>
<td>0.3206</td>
</tr>
<tr>
<td>daughter</td>
<td>0.2656</td>
</tr>
<tr>
<td>nausea</td>
<td>0.2309</td>
</tr>
<tr>
<td>female</td>
<td>0.2246</td>
</tr>
<tr>
<td>commode</td>
<td>0.2183</td>
</tr>
<tr>
<td>responded</td>
<td>0.2052</td>
</tr>
<tr>
<td>fick</td>
<td>0.2009</td>
</tr>
<tr>
<td>cco</td>
<td>0.1975</td>
</tr>
</tbody>
</table>

Table 4. Most predictive words for admission types

(a) 'Urgent' admissions

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ew</td>
<td>0.2639</td>
</tr>
<tr>
<td>er</td>
<td>0.2495</td>
</tr>
<tr>
<td>fracture</td>
<td>0.2258</td>
</tr>
<tr>
<td>fx</td>
<td>0.2248</td>
</tr>
<tr>
<td>osh</td>
<td>0.2235</td>
</tr>
<tr>
<td>b</td>
<td>0.2194</td>
</tr>
<tr>
<td>disease</td>
<td>0.2138</td>
</tr>
<tr>
<td>vertebral</td>
<td>0.2061</td>
</tr>
<tr>
<td>cabg</td>
<td>0.2029</td>
</tr>
<tr>
<td>fractures</td>
<td>0.1971</td>
</tr>
<tr>
<td>fall</td>
<td>0.1893</td>
</tr>
<tr>
<td>arteriogram</td>
<td>0.1877</td>
</tr>
</tbody>
</table>

(b) 'Elective' admissions

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>sda</td>
<td>0.8048</td>
</tr>
<tr>
<td>flap</td>
<td>0.4646</td>
</tr>
<tr>
<td>esophagectomy</td>
<td>0.4617</td>
</tr>
<tr>
<td>artery</td>
<td>0.4435</td>
</tr>
<tr>
<td>epidural</td>
<td>0.4415</td>
</tr>
<tr>
<td>valve</td>
<td>0.3845</td>
</tr>
<tr>
<td>lobectomy</td>
<td>0.3838</td>
</tr>
<tr>
<td>resection</td>
<td>0.3644</td>
</tr>
<tr>
<td>avr</td>
<td>0.3527</td>
</tr>
<tr>
<td>replacement</td>
<td>0.3324</td>
</tr>
<tr>
<td>nephrectomy</td>
<td>0.2812</td>
</tr>
<tr>
<td>whipple</td>
<td>0.2740</td>
</tr>
</tbody>
</table>

Table 5. Most predictive words for length-of-stay

(a) Short stay (0 - 1.5 days)

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ml</td>
<td>0.5295</td>
</tr>
<tr>
<td>pt</td>
<td>0.5086</td>
</tr>
<tr>
<td>to</td>
<td>0.3370</td>
</tr>
<tr>
<td>b</td>
<td>0.3403</td>
</tr>
<tr>
<td>sensitivity</td>
<td>0.2489</td>
</tr>
<tr>
<td>meq</td>
<td>0.2090</td>
</tr>
<tr>
<td>atrial</td>
<td>0.1934</td>
</tr>
<tr>
<td>tamponade</td>
<td>0.1784</td>
</tr>
<tr>
<td>valuables</td>
<td>0.1770</td>
</tr>
<tr>
<td>vomited</td>
<td>0.1738</td>
</tr>
<tr>
<td>s</td>
<td>0.1708</td>
</tr>
<tr>
<td>weaning</td>
<td>0.1676</td>
</tr>
</tbody>
</table>

(b) Medium stay (1.5 - 3.5 days)

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>followed</td>
<td>0.2014</td>
</tr>
<tr>
<td>aps</td>
<td>0.1888</td>
</tr>
<tr>
<td>lifting</td>
<td>0.1811</td>
</tr>
<tr>
<td>of</td>
<td>0.1796</td>
</tr>
<tr>
<td>device</td>
<td>0.1790</td>
</tr>
<tr>
<td>trunk</td>
<td>0.1747</td>
</tr>
<tr>
<td>available</td>
<td>0.1644</td>
</tr>
<tr>
<td>metastatic</td>
<td>0.1610</td>
</tr>
<tr>
<td>the</td>
<td>0.1603</td>
</tr>
<tr>
<td>holes</td>
<td>0.1576</td>
</tr>
<tr>
<td>this</td>
<td>0.1520</td>
</tr>
<tr>
<td>decubitus</td>
<td>0.1509</td>
</tr>
</tbody>
</table>

(c) Long stay (> 3.5 days)

<table>
<thead>
<tr>
<th>Word</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>amio</td>
<td>0.2470</td>
</tr>
<tr>
<td>mn</td>
<td>0.2393</td>
</tr>
<tr>
<td>brain</td>
<td>0.2172</td>
</tr>
<tr>
<td>decreasing</td>
<td>0.2002</td>
</tr>
<tr>
<td>fentanyl</td>
<td>0.1971</td>
</tr>
<tr>
<td>withdrawal</td>
<td>0.1933</td>
</tr>
<tr>
<td>vasospasm</td>
<td>0.1900</td>
</tr>
<tr>
<td>previously</td>
<td>0.1890</td>
</tr>
<tr>
<td>coiling</td>
<td>0.1811</td>
</tr>
<tr>
<td>exercises</td>
<td>0.1799</td>
</tr>
<tr>
<td>dobbhoff</td>
<td>0.1779</td>
</tr>
<tr>
<td>frequently</td>
<td>0.1776</td>
</tr>
</tbody>
</table>
exceedingly common words. While this step is clearly important in the treatment of “stopwords” — words that are so common they provide no additional value — here it inadvertently removes commonly recorded information. This presents a challenge for aggregating the word embeddings of a note into one single document embedding because including too many words in the aggregate statistical values (i.e., averages, maximums, and minimums) drives down the “informativeness” of the representation by adding noise to these aggregate statistics.

Further, all methods achieve high AUCs for mortality, admission type, and gender; similarly, each performs poorly for ethnicity. The highest ethnicity AUC is still 20 points lower than the worst reported AUC for the other tasks. This suggests that predicting ethnicity from notes is an inherently difficult challenge. This is largely because race, while commonly coded elsewhere, is not typically specified in the notes. Additionally, 71% of patients are white in our dataset. This class imbalance may be large enough that a “default” value may be assumed and not recorded. When ethnicity is mentioned, it is usually to denote a language barrier, e.g. “Spanish-speaking” or “required translator.”

In general, interpretability is seen as a desirable feature for machine learning, particularly in the clinical setting: doctors care not only about what decision is made, but what information is used to inform that decision. Here, BoW seems to have a natural advantage over other embedding models, as it is very easy to examine what words have the most predictive power for given tasks.

Indeed, Table 3 clearly demonstrates the interpretability of the features for predicting gender. Words such as ‘man’, ‘male’, ‘wife’, and ‘he’ directly suggest a male patient, and these are shown to have high predictive power for gender. More interestingly, we see words corresponding to gender-correlated conditions and body parts, such as ‘prostate’ for men and ‘breast’ for women. Unsurprisingly, BoW performs better than other methods on this task.

Admission type, with features shown in Table 4 is less-easily interpreted, but still provides understandable features. Words such as ‘er’ and ‘ew’ refer to the emergency room or ward, and ‘fracture’ or ‘fall’ refer to traumatic injuries, all of which reasonably suggest an urgent-care admission. Conversely, many of the predictive words for elective admissions suggest chronic conditions or planned surgical procedures (‘artery’, ‘valve’, ‘replacement’). We see that BoW also performs quite well on this task.

However, we see some differences when we examine the predictive features for the length-of-stay task in table 5. In contrast to gender or admission type, the features for length-of-stay are much more generic, seeming to have little interpretable relation to the prediction task. At the same time, we see that the LSTM achieves a higher F1 score as compared to the BoW model for this task. This suggests that BoW is interpretable for the simple token-matching tasks, but not the harder reasoning tasks. Therefore, more complex and performant models should be used for these harder tasks.

Conclusions

In this work we consider both demographic and clinical prediction tasks in order to “stress test” a variety of common note representations. We show that different representations have different strengths: while complex models can outperform simple ones on reasoning tasks, they struggle to capture seemingly “easy” information. On the other hand, simple word-matching models prove to be very effective and interpretable for tasks that are so simple that complex models tend to overlook their differences. In doing so, we motivate the need for considering multiple representations rather than adopting a one-size-fits-all approach. Finally, to promote open and reproducible research, our code is publicly available, alongside word vectors trained on a very large corpus of clinical notes.

Acknowledgments

The authors would like to thank Jen Gong for her input and suggestions. This research was funded in part by the Intel Science and Technology Center for Big Data, the National Library of Medicine Biomedical Informatics Research Training grant 2T15 LM007092-22, the National Science Foundation Graduate Research Fellowship Program under Grant No. 1122374, NIH grants U54-HG007963 and R01-EB017205, and collaborative research agreements from Philips Corporation and Wistron Corporation.
References


ICD-10 procedure codes produce transition challenges

Andrew D. Boyd MD1,8, Jianrong ‘John’ Li MSc3-6, Colleen Kenost3,7, Samir Rachid Zain3,6,7,8, Jacob Krive PhD1, Manish Mittal PhD1,2, Richard A. Satava MD9, Michael Burton MSc10, Jacob Smith3,7, Yves A. Lussier MD3-7*

1Dept of Biomedical and Health Information Sciences, University of Illinois at Chicago; 2Dept of Pharmacology, University of Illinois at Chicago; 3Center for Biomedical Informatics and Biostatistics, 4University of Arizona Health Sciences, 5Cancer Center, 6BIO5 Institute, and 7Dept of Medicine, 8Graduate Interdisciplinary Program in Statistics, The University of Arizona, Tucson, Arizona, USA; 9Dept of Surgery, University of Washington Medical Center, Seattle, Washington, USA; 10Dept of Medicine, University of Illinois at Chicago; §Joint first authorship, *Corresponding authors

Abstract

The transition of procedure coding from ICD-9-CM-Vol-3 to ICD-10-PCS has generated problems for the medical community at large resulting from the lack of clarity required to integrate two non-congruent coding systems. We hypothesized that quantifying these issues with network topology analyses offers a better understanding of the issues, and therefore we developed solutions (online tools) to empower hospital administrators and researchers to address these challenges. Five topologies were identified: “identity” (I), “class-to-subclass” (C2S), “subclass-to-class” (S2C), “convoluted(C)”, and “no mapping” (NM). The procedure codes in the 2010 Illinois Medicaid dataset (3,290 patients, 116 institutions) were categorized as C=55%, C2S=40%, I=3%, NM=2%, and S2C=1%. Majority of the problematic and ambiguous mappings (convoluted) pertained to operations in ophthalmology, cardiology, urology, gynecology, and dermatology. Finally, the algorithms were expanded into a user-friendly tool to identify problematic topologies and specify lists of procedural codes utilized by medical professionals and researchers for mitigating error-prone translations, simplifying research, and improving quality.

http://www.lussiergroup.org/transition-to-ICD10PCS

Introduction

As of October 2015, the US healthcare systems transitioned from the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) Volume 3 (ICD-9-CM-Vol3) to the ICD, Tenth Revision, Procedure Coding System (ICD-10-PCS). However, this transition was costlier and more complex than projected by the Centers for Medicare and Medicaid Services (CMS), the government agency that oversees medical coding standards and mandated the migration to the ICD-10-PCS coding platform1. To ensure data consistency at the national level during this transition, CMS and the Centers for Disease Control and Prevention (CDC) created General Equivalence Mappings (GEMs) for translating data from ICD-9-CM-Vol3 procedures to ICD-10-PCS and vice versa2. Unfortunately, GEMs have limited functionality as more than 97% of the ICD-9-CM-Vol3 codes have only ‘approximate’ matches to ICD-10-PCS codes3. Therefore, the bi-directional examination of the mappings is required to accurately match codes between the two classification systems and address the embedded imprecision. This unsolved problem led the Agency for Healthcare Research and Quality (AHRQ) to develop the MapIT tool kit – a software listing all mappings between the two sets of codes via four different tables. The very nature of the AHRQ software illustrated the complexity of translating an individual code from one system to the other4 and the lack of an international mapping standard for procedure codes motivated the authors to develop a unifying system for the community5.

A major – and often overlooked – concern with the transition of ICD-9-CM-Vol3 to ICD-10-PCS is the 18-fold increase in the number of procedure codes in ICD-10-PCS. This 18-fold increase is due to a conscious effort to favor procedural specificity (e.g., exhaustive combinations bundling as one code “approaches” and “explicit body parts”5, 6). Even though previous publications have focused on simplifying the diagnosis code transition of ICD-9-CM to ICD-10-CM7, the transition to procedure codes in ICD-10-PCS has not received the same attention, and its impact on patient care, research, and quality improvement initiatives remains largely neglected and not well understood. Perhaps, this neglect can be attributed to physicians using diagnosis codes for clinic visits, which then results in less attention on ICD-10-PCS procedure codes6. However, when it comes to hospital reimbursement, procedure codes play a major role in the continuum of care. Physician’s notes directly impact appropriate hospital billing of ICD-10-
PCS procedures as they are utilized for identifying the code most closely matching the procedure or the approach (e.g., percutaneous, open, or endoscopic variants of a procedure).\(^6\)

In addition to physician and hospital billing and reimbursement, the ICD-9-CM system is heavily utilized by public health agencies, health care professionals, and the biomedical informatics community for tracking patient interventions and outcomes. For example, researchers must be able to standardize definitions and seamlessly work between coding systems in order to work with datasets encompassing both coding systems. On the other hand, clinical trials and longitudinal studies break down if clinical endpoints, procedural interventions, and patient outcomes do not remain invariant across coding systems. Further, ICD-9-CM documentation inaccuracies have already been previously linked to decreased reliability of patient safety indicators (PSI) reporting\(^7\) and inaccurate documentation of epidemiological research studies which can greatly compromise patient safety and public health. Addressing this via the backward mapping of ICD-10-PCS codes to ICD-9-CM-Vol3 codes would allow for consistent reporting of quality outcomes\(^16\) (e.g., AHRQ PSI in the United States\(^11\) or secondary studies of readmission risks using PSI as an underlying model for investigating discharge documentation). Further, addressing inaccurate mappings would also reduce general inefficiencies in health care administration\(^12\).

We hypothesized that modeling the two systems with network topologies would allow us to better understand how to systematically standardize the two coding systems, and thus fill a major data quality gap in the biomedical informatics community. Therefore, we produced an online mapping tool that translates ICD-9-CM-Vol3 codes to the ICD-10-PCS coding system and vice versa for accurately reporting patient health metrics. This article focuses on (i) the largely neglected implications of the transition between the ICD-9-CM-Vol3 and the ICD-10-PCS coding platforms and (ii) a systematic approach to standardizing bi-directional relationships and mappings across ICD-9-CM-Vol3 and ICD-10-PCS to mitigate error-prone translations, simplify research, and improve quality of care\(^7\). The procedure codes in the 2010 Illinois Medicaid dataset are used as a validation set to demonstrate the transitional challenges between these classification systems.

**Methodology, Data, and Implementation**

We previously conducted network modeling analyses that unveiled complex, entangled, and non-reciprocal translation mappings between ICD-9-CM and ICD-10-CM billing diagnoses that we termed "convoluted"\(^5\). Our prior studies\(^13\)–\(^18\) focused exclusively on medical diagnosis codes, whereas this study focuses on procedure codes. ICD-9-CM Volume 3 procedure codes were subjected to a similar mathematical approach for the transition to ICD-10-PCS (procedure codes), where the relationship between codes are transformed into SQL code and provided in supplement materials\(^5\). The annotated translation mappings highlight codes at risk of non-straightforward and complex translations. For example, one ICD-9-CM Volume 3 procedure code can map to several ICD-10-PCS procedure codes, and those ICD-10-PCS procedure codes map backwards to other ICD-9-CM Volume 3 procedure codes in a non-reciprocal manner. No mapping occurs when GEM files do not provide any mappings from ICD-9-CM Vol3 to ICD-10-PCS in either direction.

A network map was computationally created to show the complex relationships between ICD-9-CM Volume 3 and ICD-10-PCS based on the combination of four forward and backward GEM files, followed by an computational topological motif analysis. All 3,878 ICD-9-CM Volume 3 codes were organized into five mapping categories according to the observed network topology motifs (Figure 1): “identity”(I), “class-to-subclass”(C2S), “subclass-to-class”(S2C), “convoluted”(C), and “no mapping”(NM).\(^7\) The identity transitions are mapping relationships between the two coding platforms where one ICD-9-CM Volume 3 procedure code maps to one ICD-10-PCS procedure code.\(^7\) Compared to convoluted mappings, “identity”, “class-to-subclass”, and “subclass-to-class” mappings are simpler to track, understand, and use.

Even though CMS GEM files provide forward and backward directional mapping tables from ICD-9-CM Volume 3 to ICD-10-PCS coding systems, these mappings are not necessarily reciprocal as each code could map to multiple codes. Due to the limitations of the CMS GEM files, the AHRQ developed the MapIT tool kit which provides mappings between the two sets of codes visualized in tables\(^4\). However, several knowledge gaps exist and are not addressed by these tools. Next, we depicted the indirect relationships of ICD-9-CM Volume 3 procedure code 45.16 with other procedure codes (45.14, 45.27, 44.14 and 42.24) and the difficulty of tracking such complex relationships, which is profoundly important for maintaining clinical documentation and accurately billing.

A 2010 statewide Medicaid database from Illinois containing data for 3,290 patients was utilized to evaluate the cost implications of procedures used 10 or more times. The University of Illinois at Chicago approved this research project as exempt. The procedure codes were analyzed for: “identity” (I), “class-to-subclass” (C2S), “subclass-to-
class” (S2C), “convoluted” (C), and “no mapping” (NM). The percentage of convoluted procedures is the number of procedures labeled with an ICD-9-CM procedure code that is categorized as convoluted divided by the total number of procedures in that clinical class. The most expensive categories of the ICD-9-CM procedure codes are highlighted and compared to the results from the GEM files analysis.

![Diagram](image)

**Figure 1. Motif relationship mapping between ICD-9-Vol3 and ICD-10-PCS codes.** The mapping of ICD-9-CM-Vol3 to ICD-10-PCS and backward yields complex networks that we simplified into elementary motifs represented in this figure. An ICD-9-CM-Vol3 mapping that proceeds via a convoluted motif leads to a complex interpretation of its corresponding ICD-10-PCS code(s). Due to the non-reciprocal mappings, the majority of convoluted motifs are unbounded (dashed arrows). Unbounded motifs reverberate in the network beyond the illustrated motif (out of the motif) showing unclear reciprocal mapping (A translated to B that translates back to C and C translated to D and so on). faded motifs with white background cells contain no ICD-9-CM-Vol3 codes. Each of the matrix cells comprises one or more mapping motifs that are further synthesized into five mapping categories. The four-color coding corresponds to four categories of complexity of coding as illustrated by the very similar topological motifs of each color (with the exception of the convoluted motif that groups a myriad of unbounded motifs). Blue and yellow colored cells correspond to straightforward subsumptions in a hierarchical classification system.

**Results and Discussion**

The entire bidirectional mapping network comprises the mappings of 3,878 ICD-9-CM-Vol3 codes to 99,791 ICD-10-PCS codes (Figures 1-2). A global picture of the complexity of relationships between the two coding systems (“identity”, “class-to-subclass”, and “subclass-to-class”) is shown in Figure 2A (Left network provided in detail on the web portal), whereas Figure 2B defines the levels of complexity observed in the analysis of the translation maps. Panel A shows the mapping of the ICD-9-CM-Volume 3 to ICD-10-PCS coding procedures using the science of networks. Analysis revealed that the majority of coding procedures (55%) fall under the “convoluted” category, followed by “simple” (40%), and then “no mapping” (5%). Figure 2B provides details of this complexity for each clinical category. The highest percentages of convoluted mappings are found within obstetrical procedures and operations on the eye, integumentary system, and female genital organs (Figure 2 Panel B).

While the CMS GEM file only provides simple mappings (Figure 3A), the AHRQ MapIT toolkit begins to reveal the complexity of mappings. Figure 3B shows how the application provides a complete raw listing as four CMS GEMs tables of mapping relationships between ICD-9-CM-Vol3 and ICD-10-PCS codes for code 45.16
(esophagastroduodenoscopy with closed biopsy). However, to fully understand the translation challenge, code mappings via both directions need to be viewed simultaneously through a graphical representation. Network visualization and convoluted categorization reveal the true complexity of code 45.16 (Figure 3C and 3D). To view the same concept in the AHRQ MapIT toolkit (Figure 3B), all four tables would need to be integrated together. Viewing the indirect coding relationships as single lines in a spreadsheet may result in improper translation, further leading to potentially misrepresented patient outcomes used in quality improvement metrics, patient comparison of procedures, and research analytics.

A : Bidirectional Mapping of the ICD-9-CM Procedure Codes to ICD-10-PCS

- ICD-9-CM Vol3 Code
- ICD-10-PCS Code
- Mapping

Convoluted
Simple

B: ICD-9-CM Procedure Codes Discrimination by Clinical Category

Clinical Classes (ICD-9-CM Vol. 3)
- Other Appendectomy
- Operations on Appendiceal, Nongangrenous, Glandular, Tubal and Fallopian Tubes
- Operations on Fallopian Tubes
- Operations on the Ovary, Fallopian Tube, tubes, Ovarian, Oviducts
- Operations on the Fallopian Tube, with Intraluminal Dev, Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endoscopy
- Operations on the Fallopian Tube, with Intraluminal Dev, Perc Endosco...
A. CMS file "gem_i9pcs.txt" for Vol. 3 ICD-9-CM code 45.16

B. AHRQ MapIT Toolkit, for Vol. ICD-9-CM code 45.16 with forward, backward, reverse forward, and reverse backward mapping

C. Lussierlab tool derived from GEMs for Vol. 3 ICD-9-CM code 45.16

D. Example of the solution (table) from online tool

<table>
<thead>
<tr>
<th>Submitted ICD-9-CM-</th>
<th>ICD9 TERM</th>
<th>Submitted values</th>
<th>ICD-9-CM-</th>
<th>Relationship</th>
<th>ICD-10-PCS</th>
<th>ICD10 TERM</th>
<th>Mapping Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vol3</td>
<td></td>
<td></td>
<td>Vol3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4515 0DBB0ZX 10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4516 0D958ZX 10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4516 0D968ZX 10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4516 0DB98ZX 10000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

...
In addition, complex mapping relationships present a formidable challenge for researchers. Figure 4 shows the complexity of the mapping relationships between the convoluted ICD-9-CM Volume 3 and ICD-10-PCS coding platforms. Such complexity is not always captured by commercially-available software designed to track GEM-based code mappings. For example, the ICD-9-CM Volume 3 code 45.16, which is frequently utilized statewide with billing totaling close to $1.5 million, has many complex reciprocal relationships to ICD-10-PCS codes. Further, code 45.16 has indirect links to other ICD-9-CM Volume 3 and ICD-10-PCS codes that are not immediately apparent unless all relevant mappings are diagrammed accurately. Our categorical analysis successfully produces aggregated results for ICD-9-CM Volume 3 classifications of the utmost concern. Other infrequent convoluted procedures are too costly to disregard, such as small bowel exteriorization where each procedure costs approximately $123,000. With such a great financial incentive, hospital administrators would want to track these procedures closely and adjust for the differences in mapping when accounting for cost.

Figure 5 focuses on all procedure codes. A detailed analysis of the codes from hospitals billed to Medicaid revealed a large percentage of convoluted codes (> 50%) in obstetrical, cardiovascular, and digestive system procedures, which present some of the costlier operations (Figure 5, Table 1: 3,290 patients, 3,984 procedures, 116 institutions).

Figure 5. Depiction of costs and complexity of coding per clinical specialty using the Illinois Medicaid procedural claim data set from 2010. The Medicaid reimbursement claims for each category is represented on the logarithmic scale on Y-axis as box-violin plots, whereas the degree of complexity (% of convolution) between the two coding systems is represented on the second Y-axis by a blue line. The horizontal line in each violin-box represents the median value of reimbursement claims for each category. All the procedural categories in ICD-9-CM-Vol3 exhibited more than 50% of convoluted relationships with ICD-10-PCS except the procedures in the Hemic and Lymphatic System category. The procedures in obstetrical, cardiovascular, and urinary systems represents some of the costlier operations as well exhibited highly convoluted relationships to the ICD-10-PCS coding system (>55%).

Furthermore, procedure coding in the ICD-10-PCS platform may impact coder productivity. A recent survey demonstrated a 30 to 40% reduction in productivity of professional coders impacting revenue\(^1\). Physicians and quality improvement projects should be concerned about the extensive resources required to compare procedure across the ICD-10-PCS transition.

The majority of retrospective population health studies rely on procedure details and clinical documentation accuracy of ICD-9-CM codes such as a population-based study on colorectal cancer surgery\(^14\), surgeons’ experience performing endocrine operations\(^15\), a study of causes for reoperations after back surgery\(^16\), and causes leading to low-back surgery\(^17\). Researchers may also be interested in comparisons among hospital procedures\(^18\), as well as public health topics (e.g., cause of death studies\(^19-21\)). A number of health outcomes studies rely on ICD data, including development of a clinical comorbidity index based on ICD-9 classification\(^22\), survival and changes in comorbidities after bariatric surgery\(^23\), mortality on prostate cancer risk after surgery for benign prostatic hyperplasia\(^24\), and the impact of hospital surgical volume on operative mortality for major cancer surgeries\(^25\). Lastly, there is an argument that coding variations lead to differences in reported outcomes of clinical studies, which affects the results of population-based and retrospective studies, especially in longitudinal studies that span years when both ICD versions 9 and 10 were in use\(^26\).
Table 1. High cost procedures associated to convoluted mappings between ICD. Costliest and most frequently billed ICD-9-CM procedure categories based on 2010 Illinois Medicaid reimbursement can be coded as many distinct procedures in ICD-10-CM. Arbitrary coding of convoluted mappings may lead to disputable reimbursements, under- and over-billing, as well as difficulties in measuring performance.

<table>
<thead>
<tr>
<th>ICD-9 Vol. 3 Cat.</th>
<th>Category Description</th>
<th>Total Reimbursement</th>
<th>Total Number of Procedures</th>
<th>Average Payment for Procedure</th>
<th>% of Convoluted Procedures*</th>
</tr>
</thead>
<tbody>
<tr>
<td>72 - 75</td>
<td>Obstetrical Procedures</td>
<td>$13,571,353</td>
<td>917</td>
<td>$14,800</td>
<td>100%</td>
</tr>
<tr>
<td>00 - 00</td>
<td>Procedures and Interventions (NEC)</td>
<td>$898,663</td>
<td>16</td>
<td>$56,166</td>
<td>100%</td>
</tr>
<tr>
<td>85 - 86</td>
<td>Operations on Integumentary System</td>
<td>$872,952</td>
<td>61</td>
<td>$14,311</td>
<td>100%</td>
</tr>
<tr>
<td>65 - 71</td>
<td>Operations on Female Genital Organs</td>
<td>$646,004</td>
<td>33</td>
<td>$19,576</td>
<td>100%</td>
</tr>
<tr>
<td>87 - 99</td>
<td>Misc. Diag. &amp; Therapeutic Procedures</td>
<td>$29,292,996</td>
<td>1159</td>
<td>$25,274</td>
<td>82%</td>
</tr>
<tr>
<td>35 - 39</td>
<td>Operations on Cardiovascular System</td>
<td>$8,740,992</td>
<td>182</td>
<td>$48,027</td>
<td>55%</td>
</tr>
<tr>
<td>42 - 54</td>
<td>Operations on Digestive System</td>
<td>$6,613,899</td>
<td>221</td>
<td>$29,927</td>
<td>42%</td>
</tr>
<tr>
<td>60 - 64</td>
<td>Operations on Male Genital Organs</td>
<td>$2,089,573</td>
<td>382</td>
<td>$5,470</td>
<td>0%</td>
</tr>
<tr>
<td>01 - 05</td>
<td>Operations on the Nervous System</td>
<td>$1,182,319</td>
<td>75</td>
<td>$15,764</td>
<td>0%</td>
</tr>
<tr>
<td>30 - 34</td>
<td>Operations on the Respiratory System</td>
<td>$537,174</td>
<td>11</td>
<td>$48,834</td>
<td>0%</td>
</tr>
</tbody>
</table>

* % of convoluted procedures = number of procedures with convoluted codes divided by the total number of procedures per row

The convoluted coding designations reveal potential challenges in the transition, and comparison of these complex transitions from one coding platform to another requires additional evaluation. While there is a correct notion on the part of some ICD-10-PCS researchers4 that the majority of these mapping challenges is associated with non-unique details about procedures (i.e., additional information on a smaller subset of well-defined terms), they neglect the operational, quality improvement, and research perspectives of dealing with this complexity on an ongoing basis. Coding complexity is not a one-time implementation and learning process investment; it involves continuous operations with higher complexity that carries a certain cost burden of additional staff as well as a substantial time investment of medical providers at various levels.

Regardless of these challenges, the transition to ICD-10-PCS offers valuable benefits to the US healthcare system such as reduced discharge not final billed rates and fewer denied claims5. Additionally, this transition to ICD-10 will significantly impact clinical documentation by requiring increased details for properly coding6. Evidence from the Swiss has demonstrated that through additional education and the normal “learning curve,” quality of data does improve with the ICD-10-PCS coding platform28. Researchers and healthcare professionals must understand how to interpret the data between the two coding platforms to ensure mappings are done accurately in order to guarantee consistency of results.

Limitations

A limitation of this study is the use of a single statewide Medicaid dataset. Many other insurers including Medicare may have varying costs of reimbursement than those of Illinois Medicaid. Variations in hospitals procedure coding across the country could also lead to differences in the percentage of convoluted procedures per category. Additional tools and methodologies need to be developed and published to ensure retrospective studies examining hospital ICD-9-CM Vol3 procedure codes are consistent within studies across the transition between these and ICD-10-PCS.

Conclusion

The ambiguity induced by the asymmetric, convoluted, and incomplete mappings between ICD-9-CM-Vol-3 and ICD-10-PCS has led to inaccurate integrations that prevent researchers and institutions from properly conducting studies on electronic medical records across this change-point in time. These issues are likely to lead to imprecise billing practices which directly impact metrics of patient quality and safety. As quality and safety metrics enable feedback and quality improvement initiatives, these findings may indicate a possible impact on patient care as well. Existing transition tools lack the functionality to show the complex relationships that exist between the two coding systems and fail to highlight the many challenges of this transition. Thus, we quantified these issues via network
topology analyses and developed an online tool to help professional coders, hospital administrators, patients, and researchers better navigate the transition and interpret these challenges. A complete view of the network mapping provides a clear understanding of the implications of this coding platform transition, and the mapping categories (“identity”, “class-to-subclass”, “subclass-to-class”, “convoluted”, and “no mapping”) are beneficial for conducting accurate analyses and interpretations between the two coding platforms. This online tool identifies problematic topologies of procedural codes utilized by physicians, researchers, clinics, or medical centers for mitigating error-prone translations, streamlining research, and improving quality. Without a clear understanding of these complex relationships, we may jeopardize the integrity and reliability of clinical reports and research studies, potentially compromising patient care and health outcomes.

Availability of data and materials
The data was collected from the state of Illinois Medicaid claims. The data use agreements prohibit us from depositing the raw billing data for the individual patient visits. If researchers are interested in the data, please request from the following website:

https://www.illinois.gov/hfs/MedicalProviders/cc/spwd/Pages/DataRelease.aspx

The algorithms to identify the codes as simple, convoluted, and other categories is located:
1. SQL database for the individual motifs http://lussierlab.org/publication/Motif_table_SQLcode/
2. Website: http://lussierlab.org/transition-to-ICD10PCS

Authors contribution
JJL and MB programmed the software; ADB, JJL, JK, MM, CK, JS, RAS, SRZ, and MB contributed to the analyses, figures and tables, ADB and YAL conceived and supervised the study. Every author contributed to the writing of the interpretation and discussion.

Acknowledgements
This work was supported in part by The University of Arizona Health Sciences Center for Biomedical Informatics and Biostatistics, the BIO5 Institute, and NIH/NLM (K22 LM008308).

Consent to participate
As this study included data collected as part of routine clinical care and payment, a waiver of consent was obtained through The University of Illinois at Chicago Institutional Review Board, which approved this study (2012-0773).

Competing interests
In 2014, Dr. Boyd was a paid speaker for Epic, which had no input into the design or publication of this manuscript.

References


A Network-Theoretic Analysis of Hospital Admission, Transfer, and Discharge Data

Maria Cioffi¹, Naba Mukhtar¹, Nathan C. Ryan, PhD¹, Joe J. Klobusicky, PhD²
¹Bucknell University, Lewisburg, Pennsylvania; ²Rensselaer Polytechnic Institute, Troy, New York

Abstract

Comprehending complex behavior of flow within a graph is of interest to clinicians and mathematicians alike. In this study we examine admission, discharge, and transfer data of patients within a hospital system, and process the importance of nodes through several graph metrics. One common metric, which measures population densities through a continuous time Markov process, will be compared against centrality measures, a technique more often used in social media studies. Our findings show that centrality measures capture behavior related to the topology of the network that may be missed by Markov processes. This suggests that, for determining the allocation of resources between departments of a hospital, centrality measures in some cases may prove more suitable for interpreting patient flow data. Departmental rankings and suitable instances for the application for each graph metric are provided.

Introduction

A system of communicating agents can often be successfully modeled as a network. One common model studies paths across a network, such as a web session or connecting flights between two cities. Network analysis can consider a large number of paths simultaneously. A natural question, then, is how to create satisfactory descriptors for summarizing flow through a network. For the analysis of flow between departments of a hospital, the subject of this paper, two factors play a key role for how patients transfer across the hospital network. First, there is a time component recovered from departmental transfer and holding times. This is used to describe the population dynamics across the network, which frequently involves the use of Markov models. There is also the overall topology or “shape” of the network. This is more of an issue with the connectedness of the underlying graph, and involves measures of “centrality”.

For medicine, the population dynamics of allocation in emergency departments is directly related to quality of care, patient satisfaction, and cost minimization¹⁷. Dynamic behavior of bacteria and viruses, in settings both inside and outside the hospital, also play crucial roles in understanding the spread of diseases such as malaria or HIV⁷,¹⁶,²²,²⁶. A widely used and easily interpretable model for hospital populations is the Markov process⁴,⁹,¹⁰,¹³,¹⁵. In the discrete case, known as a Markov chain, states represent a set of possible patient locations, and (directed) edges are probabilities of transferring from one state to another after one step, which may represent either a fixed time or some other well-defined sequential event (e.g., the rolling a die in a board game). For many processes involving multiple time scales, attempting to define one such step can be too coarse. This is indeed the case when surveying an entire hospital, where differences in the average length of stay are of several orders of magnitude. An alternative approach uses a continuous time Markov process. The associated network for this process assumes that transition rates between nodes are exponentially distributed, with edge weights corresponding to mean transition rates. For a large population occupying a network, densities may be approximated by a differential equation provided through the infinitesimal generator (a type of stochastic derivative) of the Markov process. One advantage to using Markov processes is the wealth of readily applicable results²⁰. For instance, in the discrete case, we have simple methods for determining probabilities of transitions after multiple steps. For both the discrete and continuous cases, we can also readily determine limiting distributions, or densities of patients after being allowed to transition for a sufficiently long period.

While Markov processes are helpful in simulations involving population estimates at each node, they do not provide a complete picture of the traffic occurring between nodes. To clarify with a simple example, consider two departments, each containing one patient. In Model A, both patients remain in their initial departments. In Model B, patients transfer between departments at a high frequency. Both models, from a Markovian analysis, will produce the same average density of one patient per department. These densities, however, do not capture interchanges which occur within the system. This consideration is relevant to our model, in which the transfer rate between two departments used in a Markovian analysis can hide important information about diverse traffic. For instance, patients entering operating units in our dataset were quite common, but an operating room holds a single patient, with a typical operating time of a few hours. Visits to neonatal intensive care units, on the other hand, occur less frequently, but may hold multiple
Table 1: Summary statistics for the ADT dataset

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients</td>
<td>46,237</td>
</tr>
<tr>
<td>Number of rooms</td>
<td>747</td>
</tr>
<tr>
<td>Number of departments</td>
<td>66</td>
</tr>
<tr>
<td>Total department transfers</td>
<td>305,048</td>
</tr>
<tr>
<td>Average number of department transfers</td>
<td>3.94</td>
</tr>
<tr>
<td>Average time during encounter</td>
<td>115.42 hours</td>
</tr>
<tr>
<td>Average time spent in a department</td>
<td>23.35 hours</td>
</tr>
<tr>
<td>Department with most admissions</td>
<td>Emergency medicine (42.19%)</td>
</tr>
<tr>
<td>Proportion of patients with at least one transfer</td>
<td>83.93%</td>
</tr>
<tr>
<td>Department with longest average stay</td>
<td>Geriatric psychology (11.84 days)</td>
</tr>
<tr>
<td>Department with shortest average stay</td>
<td>Preoperative waiting rooms (15.85 minutes)</td>
</tr>
</tbody>
</table>

infants, with holding times spanning up to several weeks. The variables in patient traffic have direct applications to epidemiology, which is highly interested in how quickly an infected agent travels across a population.

To account for nodes which may have fewer patients at a given moment, but see more traffic by virtue of how connected they are to other nodes in the network, we employ a variety of centrality metrics for networks. As these metrics consider how connected a node is to other nodes, it is fair to describe them as measurements of certain aspects of the shape of the network. These metrics were largely used in mathematical sociology in the 1970s\textsuperscript{2,11}, but have recently enjoyed a rebirth in computer science with the advent of search engines and social media\textsuperscript{12,23}. The purpose for developing different centrality methods for a given network stems from the many ways to consider a node’s importance. We may, in one case, be interested simply in the number of connections of a particular node. In another case, we might instead wish to find which nodes serve as “hubs”, or are intermediates in the shortest paths between a large number of nodes. In this paper, we have decided on degree, closeness, betweenness, and eigenvector centrality, which together provide a wide variety of methods for ranking nodes.

In \textit{Data and Methods}, we give a brief overview of each centrality metric, and how each such metric measures a node’s importance. We also review continuous time Markov processes, their connection to differential equations for describing population flows, and how to simulate solutions. Each method produces a unique ranking of nodes, which are presented in the \textit{Results} section. The implications of the difference between the various metrics are given in the \textit{Discussion} section. Finally, we summarize our findings in the \textit{Conclusion} section.

\textbf{Data and Methods}

Our dataset consists of time-stamped variables relating to ADT (admissions, discharges, and transfers), consisting of 46,247 patients admitted between December 2013 and March 2016 who permitted the use of their data for research purposes. All patients were seen at one of three hospitals (Geisinger Medical Center, Geisinger Wyoming Valley, and Geisinger Shamokin Area Community Hospital) operated by Geisinger Health System, located in northeastern and central Pennsylvania. Patients transferred through a total of 747 rooms, each of which is assigned to one of 66 departments, such as telemetry units, operating rooms, nurseries, etc. A summary of descriptive statistics is given in Table 1. For the entirety of this paper, we consider statistics for all three hospitals taken together, including transfers between hospitals. As we will see in the methods section, however, after building an appropriate network to describes patient flows, it is easy to restrict analysis to a specific hospital by simply running the necessary algorithms on a subgraph.

\textbf{Basic Network Theory Terminology}

A \textit{network} is a collection of nodes and links between them. A hospital (or more generally a healthcare system) can naturally be modeled as a network. For example, the nodes could be the departments in the hospital (or the rooms, or the doctors) and we could place a link between node \textit{A} and node \textit{B} if at some point a patient moved from room \textit{A} to room \textit{B} or vice versa. A network is said to be \textit{directed} if links go from one node to the other but not necessarily the other way. In the same example as above, we would add a link from node \textit{A} to node \textit{B} if at some point at least one
patient moved from room $A$ to room $B$. A network is said to be \textit{weighted} if a real number is attached to each link. Still with the same example, we could place a link between node $A$ and $B$ with weight $11$ to represent that $11$ patients moved from room $A$ to room $B$ at some point. See Figure 1 for an example of a weighted and directed network (all network visualizations in this paper are done using Gephi\textsuperscript{1}). This particular weighted directed network is the network that we use to represent the Geisinger Medical Center data we use in our study. Suppose a weighted directed network $S = (N, L)$ has nodes $N = n_1, \ldots, n_j$ and links $L = \ell_1, \ldots, \ell_k$. Then, since it is directed, the links are \textit{ordered} pairs $(\ell_r, \ell_s)$ corresponding to a link from node $n_r$ to node $n_s$. Since $S$ is weighted, each link $\ell_i$ has associated to it a real number called its \textit{weight} that we denote by $\text{wt}(\ell_i)$. We denote the set of nodes of $S$ by $N$ and its set of links by $L$. We use $\#N$ to denote the number of nodes.

For each node in a network, we say that its \textit{degree} is the number of links going into or out of the node. Analogously, a node’s \textit{in-degree} is the number of links going into the node and its \textit{out-degree} is the number of links leaving the node. In a weighted graph, a node’s \textit{strength} is the sum of the weights of the links either leaving or entering the node and the analogous ideas of \textit{in-strength} and \textit{out-strength} can be defined, respectively, as the sum of the weights of the links entering and leaving the node. For the network $S$ described above, then, we have the following definitions. Let $n$ be a node of $S$, then

\[
\text{ideg}(n) = \sum_{(m,n) \in L} 1, \quad \text{odeg}(n) = \sum_{(n,m) \in L} 1, \quad \text{deg}(n) = \text{ideg}(n) + \text{odeg}(n);
\]

\[
\text{istr}(n) = \sum_{\ell=(m,n) \in L} \text{wt}(\ell), \quad \text{ostr}(n) = \sum_{\ell=(n,m) \in L} \text{wt}(\ell), \quad \text{str}(n) = \text{istr}(n) + \text{ostr}(n).
\]

For two nodes in a network, a \textit{path} from one to the other is a sequence of links of the network that start at the first node and end at the second node. The \textit{shortest path} between two nodes on a weighted network is the path with the smallest sum of weights of its edges.

\section*{Markov Processes}

To understand the evolution of the total number of patients, we use a continuous time Markov process. Here, nodes (often called states in the Markovian context) are given as rooms and departments. Links between states are weighted and directed, with edge weights $a_{ij}$ corresponding to constant rates from state $i$ to state $j$. This can be converted to a probabilistic interpretation as follows. When considering a closed system, in which all patients are confined to a given collection of states, the evolution of patient density $p(t) = (p_1(t), p_2(t), \ldots, p_n(t))$, where $p_i(t)$ denotes the fraction of patients in room $i$ at time $t$, is

\[
\frac{dp_i}{dt} = \sum_{k|k \rightarrow i} a_{ki}p_k - \sum_{k|i \rightarrow k} a_{ik}p_k.
\]
For the problem of flow through a hospital, however, we must also deal with both patients who are admitted or released. Patients who leave a hospital may be represented by the inclusion of an absorbing, or “cemetery”, state. Handling admissions is more complex, and involves a patient “influx vector” $c = (c_1(t), \ldots, c_n(t), 0)$ which gives rates of admission to various rooms, with the final 0 entry corresponding to the cemetery state. For the augmented matrix $\tilde{A}$ containing an extra row and column for the absorbing state, the evolution equation for total number becomes

$$\frac{dp}{dt} = \tilde{A}p + c(t)$$

To compute coefficients, we considered rates between two states in patients per minute. Precisely, for a proportion $q(X,Y)$ of all patients in state $X$ which transfer to state $Y$, having times $t_1, \ldots, t_n$ spent in state $X$, the transition rate is given as the average

$$a(X,Y) = q(X,Y) \sum_{i=1}^{n} \frac{1}{t_i}.$$ 

Along with using the empirical rates just described, we will also consider a uniform simplification where patients, on average, spend equal time in each state. From this assumption, transition rates simplify to $a(X,Y) = nrq(X,Y)$, where $r$ is the mean time spent, where we average over all departments. Note that this cruder ranking method should be seen as comparable to the other centrality metrics, which are more concerned with connectivity, rather than actual transfer times.

For simulation of the differential equation (1), we may employ Duhamel’s formula, which provides the explicit solution

$$p(t) = e^{\tilde{A}t}p(0) + \int_{0}^{t} e^{\tilde{A}(t-s)}c(s)ds, \quad t > 0,$$

where $e^M$ denotes the matrix exponential of a matrix $M$. For our purposes, we consider initial conditions of an empty hospital ($p_0 \equiv 0$), although this methodology allows for running simulations on multiple scenarios, such as a sudden admission spike in the emergency department. With a continuous influx of patients, we expect, and in fact obtain, convergence of populations for each department, since total transition rates to the cemetery state approximately match total influx into the hospital, and all patients are eventually discharged.

**Centrality Measures**

From the perspective a network’s connectedness, it is natural to ask which nodes are “most connected” or “most central”. Interpretations of centrality vary. Freeman\textsuperscript{11} identified three such interpretations by defining notions of degree, closeness, and betweenness centrality. Bonacich\textsuperscript{2} identified a fourth centrality measure that we also consider, the so-called eigenvector centrality.

In a weighted directed network, the *in*-strength (respectively, *out*-strength) centrality of a node takes into the strength of the node and not solely its degree. We feel justified in only considering this centrality measure because we are trying to understand patient flow in a hospital system and the number of patients flowing through a node. We do mention that Opsahl-Agneessens-Skvoretz\textsuperscript{21} introduced a tuning parameter that allows one to weight the strength and degree differently, depending on what one wants to measure. Formal definitions of these centrality measures are as follows. Let $n$ be a node in a network $S = (N, L)$ with

- in-strength centrality of $n = ISC(n)$
  - $= istr(n)$,
- out-strength centrality of $n = OSC(n)$
  - $= ostr(n)$. 

(A similar model has been used\textsuperscript{14,27} in the field of self assembly in nanoscience). This can be written compactly as $\frac{dp}{dt} = Ap$ with matrix entries $a_{jk}$ for $j \neq k$, and $a_{jj} = -\sum_{k \neq j} a_{jk}p_j$. 

For the problem of flow through a hospital, however, we must also deal with both patients who are admitted or released. Patients who leave a hospital may be represented by the inclusion of an absorbing, or “cemetery”, state. Handling admissions is more complex, and involves a patient “influx vector” $c = (c_1(t), \ldots, c_n(t), 0)$ which gives rates of admission to various rooms, with the final 0 entry corresponding to the cemetery state. For the augmented matrix $\tilde{A}$ containing an extra row and column for the absorbing state, the evolution equation for total number becomes

$$\frac{dp}{dt} = \tilde{A}p + c(t)$$

To compute coefficients, we considered rates between two states in patients per minute. Precisely, for a proportion $q(X,Y)$ of all patients in state $X$ which transfer to state $Y$, having times $t_1, \ldots, t_n$ spent in state $X$, the transition rate is given as the average

$$a(X,Y) = q(X,Y) \sum_{i=1}^{n} \frac{1}{t_i}.$$ 

Along with using the empirical rates just described, we will also consider a uniform simplification where patients, on average, spend equal time in each state. From this assumption, transition rates simplify to $a(X,Y) = nrq(X,Y)$, where $r$ is the mean time spent, where we average over all departments. Note that this cruder ranking method should be seen as comparable to the other centrality metrics, which are more concerned with connectivity, rather than actual transfer times.

For simulation of the differential equation (1), we may employ Duhamel’s formula, which provides the explicit solution

$$p(t) = e^{\tilde{A}t}p(0) + \int_{0}^{t} e^{\tilde{A}(t-s)}c(s)ds, \quad t > 0,$$

where $e^M$ denotes the matrix exponential of a matrix $M$. For our purposes, we consider initial conditions of an empty hospital ($p_0 \equiv 0$), although this methodology allows for running simulations on multiple scenarios, such as a sudden admission spike in the emergency department. With a continuous influx of patients, we expect, and in fact obtain, convergence of populations for each department, since total transition rates to the cemetery state approximately match total influx into the hospital, and all patients are eventually discharged.

**Centrality Measures**

From the perspective a network’s connectedness, it is natural to ask which nodes are “most connected” or “most central”. Interpretations of centrality vary. Freeman\textsuperscript{11} identified three such interpretations by defining notions of degree, closeness, and betweenness centrality. Bonacich\textsuperscript{2} identified a fourth centrality measure that we also consider, the so-called eigenvector centrality.

In a weighted directed network, the *in*-strength (respectively, *out*-strength) centrality of a node takes into the strength of the node and not solely its degree. We feel justified in only considering this centrality measure because we are trying to understand patient flow in a hospital system and the number of patients flowing through a node. We do mention that Opsahl-Agneessens-Skvoretz\textsuperscript{21} introduced a tuning parameter that allows one to weight the strength and degree differently, depending on what one wants to measure. Formal definitions of these centrality measures are as follows. Let $n$ be a node in a network $S = (N, L)$ with

- in-strength centrality of $n = ISC(n)$
  - $= istr(n)$,
- out-strength centrality of $n = OSC(n)$
  - $= ostr(n)$.
In a weighted directed network, the *closeness centrality* of a node is the reciprocal of the node’s farness, or the sum of the node’s distance to every other node. Newman\(^{18}\) provided a suitable translation of “distance” when the graph is weighted: attach a cost (often the reciprocal of the weight) to the node in question and then that node’s distance to another node is the smallest total cost among all paths from the node in question to the destination-node. Again, a variant has been proposed\(^{21}\) that also takes into account the number of links in the least costly paths. These algorithms used Dijkstra’s famous algorithm\(^{8}\) for finding shortest paths on directed networks as their starting points. Now we provide a formal definition: Let \(n\) be a node and let \(c(m, n)\) be the cost of the link from \(m\) to \(n\), that is, in our context, the reciprocal number of patients going from node \(m\) to \(n\). A smaller cost means that the two nodes are more closely connected. Then the closeness centrality of \(n\) is defined as

\[
CC(n) = \frac{1}{\sum_{n \neq m \in E} c(m, n)}.
\]

In a weighted directed network, the *betweenness centrality* of a node is a measure of how much of a role the node has in the flow among and between the other nodes in the network. This algorithm can be modified\(^{21}\) by the introduction of a tuning parameter to allow one to take the number of links involved in the paths and not just the costs. Let \(n\) be a node in \(S = (N, L)\) and set \(\sigma(a, b)\) to be the number of least costly paths from node \(a\) to node \(b\) and let \(\sigma(a, b \mid n)\) be the number of least costly paths from node \(a\) to \(b\) that pass through node \(n\). Then the betweenness centrality is given as

\[
BC(n) = \sum_{a, b \in N} \frac{\sigma(a, b \mid n)}{\sigma(a, b)}.
\]

In a weighted directed network, the *in-eigenvector centrality* of a node is a measure of its importance determined by the importance of the nodes that link to it and those that link to them and so on. Let \(A = (a_{ij})_{1 \leq i, j \leq \#N}\) be the adjacency matrix of the network \(S = (N, L)\). That is, the entry in the \(i^{th}\) row and \(j^{th}\) column is \(w_{ij}\), the weight of the link from \(i^{th}\) node to the \(j^{th}\) node or zero if there is no such link. If the network satisfies certain technical conditions (irreducibility), then, by the Perron–Frobenius Theorem, \(A\) has \(\#N\) distinct eigenvalues, with a unique largest one \(\lambda_1\) (see Thm 1.3.5 in Kitchens\(^{28}\)). The left eigenvector \(x = \begin{pmatrix} x_1 \\ \vdots \\ x_{\#N} \end{pmatrix}\) corresponding to \(\lambda_1\), the unique solution of the matrix equation

\[
x A = \lambda_1 x,
\]

encodes the in-eigenvector centrality. That is

\[
\text{in-eigenvector centrality of } n = IEC(n) = x_i,
\]

where \(n\) is the \(i^{th}\) node in \(S\) and \(x_i\) is the \(i^{th}\) component of \(x\). To find the out-eigenvector centrality, one starts with the transpose of \(A\) and finds the right eigenvector of \(A^t\); for a node \(n\), we denote that measure by \(OEC(n)\).

Our centrality results are computed using the implementations of these centrality measures in NetworkX\(^{24}\) and custom written code available online\(^{6}\).

**Results**

We summarize the results of our calculations in a series of tables and figures. To determine a ranking of departments, we take average populations of departments estimated from the Markovian model over a period of 1000 days. All departments approach a limiting distribution for their populations. Convergence times differ for each department, but all settle by day 30. The convergence of popular departments for each of our Markov models are given in Figures 2–3.

In Table 2 we identify those nodes that are most important according to our centrality metrics described above.

**Discussion**

Above we have alluded to some of the interpretations for the various metrics. We now consider consider each of these interpretations and identify which metrics are better suited for answering appropriate questions from the perspective of hospital flows.
**Figure 2:** Left: Prevalent departments under a continuous time Markov chain, modeled as patients in a department under a constant influx of admitted patients. Note that for this graph we have used log scaling for the dependent variable. Right: Prevalent departments under a continuous time Markov chain modeled as the number of patients in a department under a constant influx of admitted patients, assuming constant transition rates between connected nodes.

![Most Populated Departments](image1)

![Most Populated Departments with Constant Transitions](image2)

**Figure 3:** Prevalent individual rooms under a continuous time Markov chain, modeled as patients in a department under a constant influx of admitted patients.

![Most Populated Rooms](image3)

**Table 2:** The top four departments by various centrality measures. We recall that ISC is in-strength centrality, OSC is out-strength-centrality, CC is closeness centrality, BC is betweenness centrality, IEC is in-eigenvector centrality and OEC is out-eigenvector centrality.

<table>
<thead>
<tr>
<th>ISC</th>
<th>OSC</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERIOP IP</td>
<td>OR IP</td>
<td>EMERGENCY MEDICINE GSACH</td>
</tr>
<tr>
<td>OR IP</td>
<td>PERIOP IP</td>
<td>EMERGENCY MEDICINE</td>
</tr>
<tr>
<td>INTRAOP IP</td>
<td>EMERGENCY MEDICINE</td>
<td>OR IP</td>
</tr>
<tr>
<td>HFAM</td>
<td>8 IP</td>
<td>INTRAOP IP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BC</th>
<th>IEC</th>
<th>OEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR IP</td>
<td>PERIOP IP</td>
<td>PERIOP IP</td>
</tr>
<tr>
<td>HFAM 8 IP</td>
<td>OR IP</td>
<td>OR IP</td>
</tr>
<tr>
<td>MED SURG - WEST WING IP GSACH</td>
<td>INTRAOP IP</td>
<td>INTRAOP IP</td>
</tr>
<tr>
<td>BP7 IP</td>
<td>HFAM 6 IP</td>
<td>IN OUT SURGERY 2 IP</td>
</tr>
</tbody>
</table>
We mention that three departments consistently appeared as the most important nodes for most metrics. Those departments were the Operating department, the Perioperating department and the Intraoperating department. In this section we will also justify why the list of other departments that appear in our results is not surprising.

We also point out that these results are for regional hospitals in Geisinger Health Systems, but in general results may be hospital dependent. For instance a community hospital or a teaching hospital might perform different kinds of procedures and have different expectations of how long a procedure might take.

Markov models  The continuous time Markov process makes predictions based on population densities. This takes into account a patient’s average length of stay and also instantaneous populations for a room. This population based ranking would heavily weight departments that can hold a large numbers of people at a particular time. Therefore, the Neonatal ICU, Nursery and Geriatric Psychology departments are those that are most likely to be populated with the largest number of people at a given time. Trivially, the most important room is the absorbing state, as every patient is eventually discharged. Our results, therefore, scale populations assuming that they have not exited the hospital.

The most classical applications of Markov processes would address issues of patient overflow. The closed form steady state analysis of this paper provides one such method for ranking populations. However, a more direct approach would be through applying “worst-case scenario” type initial conditions to the predetermined transition matrix, such as an influx of patients into an emergency department following a catastrophe.

By imposing constant transition rates to the continuous time Markov model with constant transition rates makes predictions of where the traffic through the network might bottleneck. By considering all time-steps to be the same size, accurate predictions of a node’s population are lost, but nodes that score highly in this metric are considered to be, on average, those which do not clear people out as quickly as they take them in. That the top three nodes according to this metric are the ones consistently identified, as we will see below, by the other “topological” metrics is not surprising.

In- and out-strength centrality  In-strength centrality measures the number of patients who have flowed into the node considered over a dataset’s time range. The top three rooms for in-strength centrality are the Perioperative department, the Operating department and the Intraoperative department. These are not surprising because these departments serve as a hub, accepting multiple patients from a variety of departments. In general, in strength centrality measure a department’s “receptivity” to other departments taken as a whole.

Similarly, out-strength centrality measures the number of patients flowing out of the department. A node with a few heavy-weighted links out of it or one with many light-weighted links out of it would have a large out-strength centrality score. Again, three of the top four are the Operating department and the Intraoperative department. The other one in the top four is Emergency Medicine, which is reasonable, as an Emergency Medicine department often admits patients to other departments.

Closeness centrality  Department $A$ and Department $B$ (rooms, beds, etc.) are close, in the sense of closeness centrality, if many patients transfer between departments $A$ to $B$. A department with a large closeness centrality score is then one which is close many departments. Our results show that two Emergency departments are highly ranked. In general, closeness centrality may help find those departments in which patients will eventually travel to many other departments. This differs slightly from in and out strength centrality, in that closeness centrality is not local in the sense that the ease of traveling through a graph considers paths of length larger than 1.

Betweenness centrality  A department $A$ is between two departments $X$ and $Y$ if, among all ways to get from department $X$ to department $Y$, a large number of paths go through $A$. So, $A$ would have a large betweenness centrality score if $A$ is highly between all pairs of nodes. That is, a large number of patients flow through node $A$ and $A$ is the most efficient (in terms of not diverting population away from their main path) of all nodes. The rooms identified in the results play exactly that role in the hospital.

There is a possible connection with a betweenness score and the spread of infections, such as C Diff., in a hospital. A room which has not been properly disinfected may be more damaging if it is a common room that many patients pass
through. This may be contrasted with a room that at any time may contain many patients, but is isolated from other departments.

**In- and out-eigenvector centrality** A department with high in-eigenvector centrality is one that linked to many other nodes which are also highly connected. The three departments that score most highly in this metric, namely the Operating, Perioperating, Intraoperating departments, are ones with have a lot of long and highly weighted paths that terminate there. It might seem, then, that the nodes with largest in-eigenvector centrality are the same as those that score highest using the Markov chain metric. The difference, though, is that in the Markov model a patient can stay in a department over two time steps, making it a good measure of how where there might be bottlenecks. In the in-eigenvector centrality, we are measuring which departments receive a lot of patients from departments with many patients, which themselves receive many patients, and so forth.

A department with high out-vector centrality is one that links to other departments with high out-vector centrality and which are themselves linked to departments with high out-vector centrality, etc. Departments with large in-eigenvector centrality, like those with large Markov Chain steady states, represent bottlenecks. Whereas important departments from the Markov Chain encode some coarse information about time (e.g., a patient might stay in a node for several time steps), the in-eigenvector centrality does not. The departments with large in-eigenvector centrality, then, identify departments that are bottlenecks because of the number of steps involved, ignoring how long the steps might take. Identifying departments that score high in this metric would allow administrators another opportunity to disperse bottlenecks.

**Conclusion**

In this work, we have applied several graph theoretic techniques for determining the relative importance of hospital departments. Each technique carries a unique interpretation, which we found to align with intuition regarding traffic in hospitals. For instance, emergency departments are highly prevalent in closeness centrality, which roughly measures ease of access to other departments, and nurseries are highly ranked in a Markovian analysis, due to average prolonged waiting times which increase average population. We find, from the variety of prevalent nodes arrived at through different graphical centrality measures, that a full description of a patient flow in a hospital may require multiple techniques. This is especially true when considering more general networks, in which a large number of scales may occur, both in time and state (e.g. rooms versus departments as nodes in a graph).

While a common theme in operations research involves Markovian analysis to answer questions regarding population, we hope that a deeper investigation of alternative centrality measures can reveal information pertaining to deeper questions of queueing theory such as bottlenecking or rates or communicability between departments, either as a result of protocol or practitioner error. We also hope in future work to address how centrality measures may relate to the optimized care regarding histories of an individual patient post-surgery. Specifically, we will look to find associations between hospital acquired infections and certain departmental pathways which are weighted by importance obtained through the methods discussed here.

**Acknowledgements**

All four authors were supported by funds from the Bucknell/Geisinger Research Initiative. We would also like to thank Jason Brown for providing us with data and answering data-related questions, Mihai Banciu for helpful discussions and Paul Griffin for helpful feedback.

**References**

Automatically Linking Registered Clinical Trials to their Published Results with Deep Highway Networks

Travis R. Goodwin, MS1, Michael A. Skinner, MD1,2, Sanda M. Harabagiu, PhD1

1The University of Texas at Dallas, Department of Computer Science, Richardson, TX; 2The University of Texas Southwestern Medical Center, Department of Surgery, Dallas, TX

Abstract

As medical science continues to advance, health care professionals and researchers are increasingly turning to clinical trials to obtain evidence supporting best-practice treatment options. While clinical trial registries such as ClinicalTrials.gov aim to facilitate these needs, it has been shown that many trials in the registry do not contain links to their published results. To address this problem, we present NCT Link, a system for automatically linking registered clinical trials to published MEDLINE articles reporting their results. NCT Link incorporates state-of-the-art deep learning and information retrieval techniques by automatically learning a Deep Highway Network (DHN) that estimates the likelihood that a MEDLINE article reports the results of a clinical trial. Our experimental results indicate that NCT Link obtains 30%-58% improved performance over previously reported automatic systems, suggesting that NCT Link could become a valuable tool for health care providers seeking to deliver best-practice medical care informed by evidence of clinical trials as well as (a) researchers investigating selective publication and reporting of clinical trial outcomes, and (b) study designers seeking to avoid unnecessary duplication of research efforts.

Introduction

Seeking to deliver best-practice medical care, clinicians increasingly rely on information provided by published guidelines and systematic reviews. However, recent analyses have estimated that less than 15 percent of major medical guidelines are supported by high-quality evidence.1,2 To bridge this gap, health care professionals are increasingly turning to evidence from clinical trials to help evaluate different treatment options.3 To provide more convenient access to clinical trials for persons with serious medical conditions and to make the results of clinical trial more available to health care providers, the United States Congress mandated in 1997 the development of the online trial registry ClinicalTrials.gov. In 2007, in accordance with the increasing role of evidence-based medicine, the mandate was expanded by requiring the timely inclusion of clinical trial results within the registry for all sponsors of non-phase-1 human trials seeking FDA approval for a new device or drug.4 Moreover, to further increase the availability of study information to patients, physicians, and investigators, the International Committee of Medical Journal Editors (ICMJE) mandated the registration of trials before considering publication of trial results.3

Unfortunately, despite the numerous policies intended to improve the timely accessibility of clinical trial results to clinicians, there remain several barriers hindering effective use of these important data. First, sponsors and investigators have inconsistently complied with the requirement to update the registry with trial results. In an evaluation of eligible human studies registered at ClinicalTrials.gov, Anderson et al. (2015)5 found that only 13.4% of the trials reported summary results within 12 months of study completion, and only 38.3% of the registered studies reported any results at any time. Moreover, once trial results are published in peer-reviewed literature, the article citation is only provided to the ClinicalTrials.gov registry in about 23%-31% of cases.6,7 When registered trials with no reported publications were manually reviewed, both Ross et al. (2009)6 as well as Huser and Cimino (2013)7 were able to find relevant MEDLINE articles for 31%-45% of reviewed clinical trials. Finally, despite the ICMJE recommendation that pertinent publications of trial results should contain a specific citation of the trial registry number to allow simple retrieval of the article with a MEDLINE search,8 this information is included in only about 7% of articles presenting trial results.

Recently, Bashir et al. (2017)9 conducted a systematic review of studies examining “links” between registered clinical trials and the publications reporting their results and found that 83% of studies required some level of manual (i.e., human) analysis (with 19% involving strictly manual analyses, 64% involving both manual and automatic analyses and 17% involving automatic analyses). They also observed that despite the increasing pressures from journal editors
to provide information about any clinical trials associated with a publication, the number of articles amenable to being automatically linked to the clinical trials they report has not increased over time. Finally, they found that automatic methods were only able to identify a median of 23% of articles reporting the results of registered trials, leading them to conclude that identifying publications reporting the results of a clinical trial remains an arduous, manual task. Clearly, there is a need for the creation of robust methods to automatically link clinical trials with their results in the medical literature.

In this paper, we present NCT Link, a system for automatically linking registered clinical trials to articles reporting their results. NCT Link incorporates state-of-the-art deep learning techniques through a specialized Deep Highway Network (DHN) designed to determine the likelihood that a link exists between an article and a clinical trial by considering the variety of information about the article, the trial, and the relationships (if any) between them. Our experiments demonstrate that NCT Link provides a 30%-58% improvement over the automatic methods surveyed in Bashir et al. (2017); consequently, we believe that NCT Link will provide a valuable tool for health care providers seeking to obtain timely access to the publications reporting the results of clinical trials. Moreover, we surmise that NCT Link may also benefit (a) researchers investigating selective publication and reporting of clinical trial outcomes and (b) study designers aiming to avoid unnecessary duplication of research efforts.

**Linking Registered Clinical Trials to their Published Results**

In previous studies examining links between registered clinical trials and published articles, investigators have described at least three ways that a published article may be considered linked to a clinical trial. For example, an article may (1) relate in some way to the trial, e.g., by providing supporting evidence for the intervention or highlighting limitations of previous, related studies; (2) be cited in the summary or official description of the trial; or (3) report the results of the trial. In this work, we focus exclusively on the third type of link: articles which report the results of a clinical trial; consequently, we consider a publication to be linked to a clinical trial if and only if it reports the results of the trial. Moreover, as in Huser and Cimino (2013), we only consider links between clinical trials registered to ClinicalTrials.gov and published articles indexed by MEDLINE.

**Figure 1:** Architecture of NCT Link.

NCT Link, illustrated in Figure 1, operates in five steps:

1. **Trial Search:** given an NCT ID, the (meta)data associated with the trial, denoted as $t$, is obtained from the registry at ClinicalTrials.gov;
2. **Article Search:** the information in $t$ is used to obtain a subset of potentially-linked articles (along with their metadata), denoted as $A = a_1, a_2, \ldots, a_L$, using a specialized local MEDLINE index (where $L$ is the maximum number of articles considered by NCT Link);
3. **L2R: Feature Extraction:** each article $a_i \in A$ retrieved for $t$ is associated with a feature vector $v_i$ encoding a number of complex features characterizing information about $t$, $a_i$, and the relationship between them;
4. **L2R: Deep Highway Network:** a Deep Highway Network (DHN) is used to infer a score $s_i$ for each article $a_i \in A$ quantifying the likelihood that $a_i$ should be linked to (i.e. reports the results of) $t$;

*NCT Link is named after the Clinical Trial identifier, NCT ID, used by ClinicalTrials.gov.
5. **L2R: Ranking**: the score \( s_i \) associated with each article \( a_i \) is used to produce a ranked list of published articles such that the rank of each article corresponds to the likelihood that it reports the results of \( t \).

In the remainder of this section, we provide a detailed description each of the five steps listed above.

**A. Searching Clinical Trials**

NCT Link operates on an NCT ID specified by the user. The NCT ID is used to obtain all the (meta)data stored in the ClinicalTrials.gov registry for the given trial. While the National Library of Medicine (NLM) provides an online interface for programmatically obtaining data about a clinical trial specified by an NCT ID, to reduce the burden on the NLM’s servers potentially imposed by our experiments, we instead created and used our own offline index of all clinical trials registered on ClinicalTrials.gov.

**Representing clinical trials.** Due to the significant variation in the amount of data associated with clinical trials, NCT Link considers only eight key aspects of each clinical trial: (1) the set of investigators* associated with the trial, (2) the set of unique institutions associated with any investigators, (3) the NCT ID of the trial, (4) the set of interventions studied in the trial, (5) the set of conditions studied in the trial, (6) the set of keywords provided to the registry, (7) the set of Medical Subject Headings (MeSH) terms provided to the registry, and (8) the completion date of the trial†. In the remainder of this paper, we use \( t \) to simultaneously refer to a clinical trial as well as all eight aspects of information associated with the trial.

**B. Searching MEDLINE Articles**

Because MEDLINE contains over 14 million articles, rather than applying the learning-to-rank component to process and score every article in MEDLINE, we first obtain a smaller, “high-recall” sub-set of candidate MEDLINE articles that are likely to report the results of \( t \). In this section, we describe the MEDLINE searching strategy used for both (1) obtaining this high-recall set of candidate MEDLINE articles as well as (2) feature extraction (described later).

**Indexing MEDLINE articles.** To search MEDLINE, NCT Link incorporates its own internal, offline index of every article in MEDLINE. This index encodes eight fields (i.e., metadata attributes) for each article in MEDLINE: (1) the authors‡ of the article (if any), (2) the investigators‡ of the article (if any), (3) the PubMed identifier (PMID) associated with the article, (4) the accession numbers (e.g. NCT IDs) of any ClinicalTrials.gov entries in the list of “DataBanks” associated with the article, (5) the full unstructured text of the abstract§, (6) the title of the article, (7) any MeSH terms associated with the article, and (8) the publication date of the article.

**Query formulation.** A clinical trial \( t \) is represented by a disjunctive Boolean query in which each aspect corresponds to a clause. Each clause, in turn, is represented by a disjunction of natural language terms (or phrases) encoding the values (e.g., investigators, conditions, etc.) associated with that aspect. Interventions, conditions, and keywords are expanded using synonyms provided by the Unified Medical Language System. To account for variations in the way affiliations were expressed, each affiliation was represented by a sequence of “partial locations” by splitting the text of the affiliation (e.g., “University of California, San Francisco”) on occurrences of commas (e.g., “University of California” and “San Francisco”). Likewise, due to differences in how authors and investigators are reported to MEDLINE by various journals, each author/investigator is represented by a sieve consisting of four queries, each less specific than the previous: (1) first name, middle initial, and last name, (2) first initial, middle initial and last name, (3) first name and last name, and (4) the first initial and last name. To account for the progressive loss of specificity, we associated each query with a weight of 1.0, 0.5, 0.3, and 0.2, respectively, which multiplicatively affects the score (described below) of any article retrieved for the query. The clause associated with each aspect is restricted to the set of semantically related fields illustrated in Figure 2.

**Scoring MEDLINE articles.** When searching our internal MEDLINE index, candidate articles are retrieved (i.e.

---

*Investigators are represented in the registry through three structured fields indicating the investigator’s first, middle, and last names.
†If the completion date is unspecified, or if the trial is not yet complete, the start date of the trial is used.
‡In MEDLINE, authors and investigators are encoded by structured fields corresponding to their first and last names as well as their initials.
§For structured abstracts, the content of all sections was combined to create a single unstructured passage of text.
selecting and scoring using the BM25\textsuperscript{12} relevance model.\textsuperscript{*} This allows the high-recall set of candidate articles \( A \) to be defined as the top ranking retrieved articles \( a_1, a_2, \cdots, a_L \), where \( L \) is the number of candidate MEDLINE articles considered by NCT Link. Conceptually, \( L \) acts as an upper bound on the number of articles the user of the system might be interested in examining. In our experiments, to ensure thorough evaluations, we used \( L = 2,000 \). However, for general use, a smaller value of \( L \) should be sufficient, e.g., \( L = 100 \).

C. Learning-to-Rank (L2R)

Given a clinical trial \( t \), and a set of articles \( A = a_1, a_2, \cdots, a_L \), the learning-to-rank module is responsible for (1) extracting features encoding the relationship between each article \( a_i \) and the clinical trial \( t \); (2) training (or using) state-of-the-art deep learning methods – a Deep Highway Network – to score each article based on the likelihood that it reports the results of \( t \); and (3) produce a ranked list of MEDLINE articles sorted by their scores.

D. Feature Extraction from MEDLINE Articles and Clinical Trials

Determining whether a link exists between an article \( a_i \) and a clinical trial \( t \) requires considering a large variety of information which varies from trial to trial and article to article. For this reason, deciding whether an article \( a_i \in A \) reports the results of \( t \) requires access to a rich set of features. When extracting features, we consider (1) the eight aspects of clinical trial \( t \) (described in Searching Clinical Trials), (2) the eight fields associated with article \( a_i \) (described in Searching MEDLINE Articles), and (3) the mapping between aspects of \( t \) and the corresponding fields of \( a_i \) illustrated in Figure 2. Table 1 lists all the features extracted for each article \( a_i \) retrieved for trial \( t \) as well as the domain (i.e. number and type of values) of each feature, where \( \mathbb{N} \) denotes the set of natural numbers, \( \mathbb{R} \) denotes the set of real numbers, and the exponent indicates the number of values (e.g. \( \mathbb{R}^5 \) corresponds to five distinct real numbers).

As shown in Table 1, three types of features are extracted: (1) trial features \( (F_1 - F_3) \), encoding information about \( t \) which is independent of \( a_i \); (2) dynamic features \( (F_5 - F_{29}) \), encoding information about the relationship between \( a_i \) and \( t \); and (3) article features \( (F_{30} - F_{33}) \), encoding information about \( a_i \) which is independent of \( t \). Features \( F_1 - F_3 \) allow the model to account for that fact that the more investigators, interventions, or conditions associated with \( t \), the more likely it is that an article will have an investigator, intervention, or condition in common with \( t \). Features \( F_5 - F_{29} \) adapt four commonly used relevance models to act as similarity measures between an aspect of \( t \) and an article \( a_i \). Specifically, we used: (1) the Best Match 25\textsuperscript{12} (BM25), (2) Dirichlet-Smoothed language model probability\textsuperscript{13} (LMD), (3) Axiomatic relevance\textsuperscript{14} (F2Exp), and (4) Divergence from Independence\textsuperscript{15} (DFI). To account for the significant variance in the number of investigators, as well as the prevalence of common names, conditions, or interventions, \( F_{18} - F_{29} \) measure five statistics capturing the similarity between each investigator, condition, or

\textsuperscript{*}To reduce the impact of abstract length on the ranking of candidate MEDLINE articles, we specified the BM25 document-length normalization term as \( k_1 = 0.25 \) rather than standard value of 0.75,\textsuperscript{12}
Table 1: Features extracted for each article $a_i$ retrieved for trial $t$.

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$ number of investigators in $t$</td>
<td>N</td>
</tr>
<tr>
<td>$F_2$ number of interventions in $t$</td>
<td>N</td>
</tr>
<tr>
<td>$F_3$ number of conditions in $t$</td>
<td>N</td>
</tr>
<tr>
<td>$F_4$ completion date of $t$</td>
<td>N</td>
</tr>
<tr>
<td>$F_5$ days elapsed between the publication date of $a_i$ and the completion date of $t$</td>
<td>N</td>
</tr>
<tr>
<td>$F_6$ BM25 from the NCT ID of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_7$ F2Exp from the NCT ID of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_8$ DFI from the NCT ID of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_9$ LMD from the NCT ID of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{10}$ BM25 from all keywords of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{11}$ F2Exp from all keywords of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{12}$ DFI from all keywords of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{13}$ LMD from all keywords of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{14}$ BM25 from all MeSH terms of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{15}$ F2Exp from all MeSH terms of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{16}$ DFI from all MeSH terms of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{17}$ LMD from all MeSH terms of $t$ to $a_i$</td>
<td>R</td>
</tr>
<tr>
<td>$F_{18}$ BM25 statistics from each investigator in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{19}$ F2Exp statistics from each investigator in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{20}$ DFI statistics from each investigator in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{21}$ LMD statistics from each investigator in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{22}$ BM25 statistics from each intervention in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{23}$ F2Exp statistics from each intervention in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{24}$ DFI statistics from each intervention in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{25}$ LMD statistics from each intervention in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{26}$ BM25 statistics from each condition in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{27}$ F2Exp statistics from each condition in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{28}$ DFI statistics from each condition in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{29}$ LMD statistics from each condition in $t$ to $a_i$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>$F_{30}$ number of authors in $a_i$</td>
<td>N</td>
</tr>
<tr>
<td>$F_{31}$ number of investigators in $a_i$</td>
<td>N</td>
</tr>
<tr>
<td>$F_{32}$ publication date of $a_i$</td>
<td>N</td>
</tr>
<tr>
<td>$F_{33}$ publication type(s) of $a_i$</td>
<td>${0,1}^{38}$</td>
</tr>
</tbody>
</table>

intervention in $t$ and $a_i$, namely, the mean, minimum, maximum, variance, and sum. Feature $F_{33}$ encodes the MeSH publication type(s) associated with $a_i$ (in our experiments, we encountered only 38 different types of publications). The values features $F_1$ - $F_{33}$ are concatenated together to form a single vector $v_i$, allowing the Deep Highway Network to consider and combine a variety of different interactions between the aspects of $t$ and the fields of article $a_i$.

E. The Deep Highway Network

Owing to the lack of clear and exact criteria for determining whether a link exists between $t$ and $a_i$, we were interested in applying deep learning techniques to automatically learn contextual high-level and expressive “meta”-features by combining the elements of $v_i$. However, a common problem when designing deep learning networks is that the there are no clear criteria or guidelines for deciding the number (and configuration) of internal or “deep” layers in the network. Fortunately, by taking advantage of recent advances in deep structure learning, we were able to define a deep neural network which automatically tunes the number of internal layers used. Specifically, we implemented a Deep Highway Network\(^{10}\) (DHN). Unlike traditional deep networks, in which information flows through each layer of the network sequentially, DHNs allow information to “skip” layers in the network by traveling along a so-called “information highway”. Thus, in a DHN, the number of specified internal layers acts as an upper bound on the number of layers used by the model. In fact, the information highway allows DHNs to be constructed with hundreds of intermediate layers – for example DHNs with over 1,000 intermediate layers have been reported.\(^{10}\) The DHN we have implemented within NCT Link considers a maximum of $10^6$ internal layers and is illustrated in Figure 3.

As shown, the main component of each intermediate layer, $l$, is a Rectified Linear Unit (ReLU),\(^{17}\) i.e.,

$$x_{l+1} = \text{ReLU}(x_l) = \max(x_l, 0),$$

where $x_l$ indicates the output of layer $l$ and 0 denotes a zero-vector. In the DHN, each ReLU layer is augmented with a highway mechanism composed of two gates: (1) a transform gate, $T \in [0,1]$, which learns a weight that is applied to the output of the ReLU, and (2) a carry gate, $C = 1 - T$, which learns whether to skip, apply, or partially apply

\(^{*}\) Additionally, and perhaps more importantly, the information highway allows the gradient to directly influence each layer during back propagation, effectively eliminating the vanishing gradient problem and allowing very deep networks to be trained.

\(^{10}\) We also experiment with 100 internal layers and observed no discernible change in performance.
the ReLU in the layer to $x_l$. Thus, the highway mechanism enables the network to learn how many (and which) layers should be applied. Formally, we define each layer in our DHN as follows:

$$T(x_l) = \sigma(x_l \cdot W_T^{(l)})$$
$$x_{l+1} = T(x_l) \cdot \text{ReLU}(W_H^{(l)} \cdot x_l) + (1 - T(x_l)) \cdot x_l$$

where $W_T^{(l)}, W_H^{(l)} \in \theta$ correspond to the learned weights of the transform gate and ReLU used in layer $l$. Figure 4 illustrates the difference between a standard ReLU layer and a ReLU layer incorporating a highway mechanism.

Figure 3: Architecture of the Highway Network used in NCT Link.

The ReLU in the layer to $x_l$. Thus, the highway mechanism enables the network to learn how many (and which) layers should be applied. Formally, we define each layer in our DHN as follows:

$$T(x_l) = \sigma(x_l \cdot W_T^{(l)})$$
$$x_{l+1} = T(x_l) \cdot \text{ReLU}(W_H^{(l)} \cdot x_l) + (1 - T(x_l)) \cdot x_l$$

where $W_T^{(l)}, W_H^{(l)} \in \theta$ correspond to the learned weights of the transform gate and ReLU used in layer $l$. Figure 4 illustrates the difference between a standard ReLU layer and a ReLU layer incorporating a highway mechanism.

To produce the score $s_i \in [0, 1]$ (i.e., the likelihood that $a_i$ reports the results of $t$) associated with $a_i$, the output of the final highway layer is projected down to a single element projected into the range $[0, 1]$ by a sigmoid layer:

$$s_i = \sigma(W_s \cdot x_F)$$

where $\sigma(x) = e^x / (e^x + 1)$ is the logistic sigmoid function, $W_s \in \theta$ corresponds to the learned weights of the final projection layer, and $x_F$ indicates the output of the final ReLU layer.

Training the Deep Highway Network. Training the DHN was achieved by finding the parameters $\theta$ most likely to predict the correct score $s_i$ for every article $a_i \in A$ retrieved for each clinical trial $t$ in the training set $T$ (details about the training set and relevance judgments used in our experiments are provided in the Experiments section). Formally, let $y_{t,i}$ indicate the relevance judgment of article $a_i$ with respect to trial $t$ such that $y_{t,i} = 1$ if $a_i$ is relevant to (i.e., reports the results of) $t$ and $y_{t,i} = 0$, otherwise. We minimize the entropy loss between the score $s_i$ assigned by the DHN and the relevance judgment $y_{t,i}$:

$$\mathcal{L}(\theta) = \sum_{(t, L) \in T} \left( \sum_{i=1}^{L} s_i \log y_{t,i} + (1 - s_i) \log (1 - y_{t,i}) \right)$$

The model was trained using Adaptive Moment Estimation\(^\text{18}\) (ADAM) (using the default initial learning rate $\eta = 0.001$).

F. Ranking MEDLINE Articles

After producing a score $s_i$ for each article $a_i \in A$ retrieved for trial $t$, the final ranked list of articles is produced by sorting the articles $a_1, a_2, \cdots, a_L$ in descending order according to their scores $s_1, s_2, \cdots, s_L$.

Experiments

Each clinical trial in ClinicalTrials.gov was manually registered by a Study Record Manager (SRM) and may be associated with two types of publications corresponding to distinct fields in the registry: (1) "related articles", articles the
SRM deemed related to the trial (typically references) and (2) “result articles”, articles the SRM indicated as reporting the results of the trial. To evaluate NCT Link, we randomly selected 500 clinical trials which were each associated with at least one “result article” in the registry. In our experiments, we used a standard 3:1:1 split for training, development, and testing. Relevance judgments for all 500 trials were automatically produced using the “result articles” encoded for each trial. Specifically, for each trial \( t \), we assigned a judgment of RELEVANT to all MEDLINE articles listed as “result articles” for \( t \). We considered two strategies for producing IRRELEVANT judgments. Initially, we applied the Closed World Assumption (CWA) by judging every MEDLINE article not explicitly listed in the “result articles” of \( t \) as IRRELEVANT to \( t \). We refer to this judgment strategy as CLOSED.

However, it has been shown that the SRM of a clinical trial does not always update the registry as new articles are published. Under the CWA, these articles would be mistakenly labeled IRRELEVANT. To account for this, we considered a secondary judgment strategy intended to minimize the likelihood of assigning an IRRELEVANT judgment to a MEDLINE article that may report the results of a trial despite not being included in the “result articles” of the trial. Formally, for each trial \( t \) we obtained a list \( A \) of 3,000 MEDLINE articles using the search strategy described in the Searching MEDLINE Articles section (without applying the learning-to-rank component of NCT Link). We determined the set of IRRELEVANT articles for \( t \) as: (1) articles which were not listed in the “result articles” of \( t \) but were listed in the “result articles” of any other trial in the registry, (2) 10 randomly selected articles between ranks 10 and 100, (3) 10 randomly selected articles between ranks 1000 and 2000, (3) 10 randomly selected articles between ranks 2000 and 3000, and (4) 10 randomly selected articles from MEDLINE not in \( A \). We refer to this second judgment strategy as OPEN. We report the performance of NCT Link when trained using the OPEN strategy and evaluated using both strategies.*

Due to the paucity of published automatic systems for linking clinical trials to their results in the literature, we measured the performance of NCT Link against two baseline systems as well as four alternative configurations of NCT Link:

1. **Exact Match**: A system in which an article is considered to be linked to a clinical trial if it specifically mentions the NCT ID of the trial in its abstract or metadata – this is an extension of the automatic approach described by Bashir et al. (2017) which considers only metadata.
2. **IR:BM25**: An information retrieval (IR) system which represents all aspects of the clinical trial as a single disjunctive Boolean query relying on the BM25 similarity function.
3. **NCT Link: BM25**: A configuration of NCT Link in which no learning-to-rank is performed; that is, the system returns the ranked list of candidate articles described in the Searching MEDLINE Articles section.
4. **NCT Link: Linear Regression**: A configuration of NCT Link that replaces the Deep Highway Network (DHN) with a linear regression model to determine article scores.
5. **NCT Link: Random Forests**: A configuration of NCT Link that replaces the DHN with a Random Forest model to determine article scores.
6. **NCT Link: Gradient Boosting**: A configuration of NCT Link that replaces the DHN with a Gradient Boosting model to determine article scores; Gradient Boosting can be viewed as modern extension to Random Forests that incorporates boosting rather than bagging to combine the scores predicted by each decision tree in the forest.

**Quality Metrics.** We measured the quality of ranked MEDLINE articles produced by all systems using standard metrics for evaluating the performance of information retrieval systems. Formally, let \( X \) indicate the test set, consisting of pairs of a clinical trial, \( t \), and the final ranked list of \( L \) articles produced for \( t \), \( B_{1:L} \). To measure the overall ranking produced by each system, we measured the Mean Average Precision (MAP):

\[
\text{MAP}(X) = \sum_{(t,b_{1:L}) \in X} \text{AP}(B_{1:L};t) ; \quad \text{AP}(B_{1:L};t) = \frac{1}{L} \sum_{k=1}^{L} \left( \text{Rel}(b_k; t) \cdot \text{Rel}(b_k; t) \right) ; \quad \text{P}(B_{1:k};t) = \frac{\sum_{k=1}^{K} \text{Rel}(b_k; t)}{\text{Num Rel}(t)}
\]

where \( \text{AP}(B_{1:L}, t) \) indicates the Average Precision of \( B_{1:L} \) with respect to \( t \), \( \text{P}(B_{1:k};t) \) represents the precision of the top-K ranked articles retrieved for trial \( t \), \( \text{Rel}(b_k; t) \) is an indicator function returning the value 1 if article \( b_k \) was judged as RELEVANT for trial \( t \) and returning 0, otherwise, and \( \text{Num Rel}(t) \) returns the number of articles judged RELEVANT for \( t \). In addition to the MAP, we report the Mean Reciprocal Rank (MRR) which is the average of the multiplicative inverse of the rank of the first relevant article produced for each trial. The MRR captures how many

---

*We found that training with the CLOSED strategy degraded performance on the test set in all cases.
irrelevant reports are ranked, on average, above the first relevant article for each trial. We also report the average precision over all clinical trials at three different ranks: the R-Precision (R-Prec) which is the precision of the first \( R \)-ranked articles, where for each trial \( t \), \( R = \text{Num Rel} (t) \); the Precision of the top-five ranked articles (P@5) and the Precision of the top-ten ranked articles (P@10).

Table 2: Quality of ranked list of MEDLINE articles retrieved for each clinical trial.

<table>
<thead>
<tr>
<th>System</th>
<th>CLOSED Strategy</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAP</td>
<td>MRR</td>
<td>R-Prec</td>
<td>P@5</td>
<td>P@10</td>
<td>MAP</td>
<td>MRR</td>
<td>R-Prec</td>
<td>P@5</td>
<td>P@10</td>
<td></td>
</tr>
<tr>
<td>Exact Match</td>
<td>0.001</td>
<td>0.001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.000</td>
<td>0.236</td>
<td>0.260</td>
<td>0.2203</td>
<td>0.0620</td>
<td>0.031</td>
<td></td>
</tr>
<tr>
<td>IR: BM25</td>
<td>0.011</td>
<td>0.016</td>
<td>0.0032</td>
<td>0.0040</td>
<td>0.003</td>
<td>0.258</td>
<td>0.294</td>
<td>0.1793</td>
<td>0.1220</td>
<td>0.095</td>
<td></td>
</tr>
<tr>
<td>NCT Link: BM25</td>
<td>0.017</td>
<td>0.021</td>
<td>0.002</td>
<td>0.004</td>
<td>0.004</td>
<td>0.586</td>
<td>0.610</td>
<td>0.549</td>
<td>0.210</td>
<td>0.115</td>
<td></td>
</tr>
<tr>
<td>NCT Link: Linear Regression</td>
<td>0.269</td>
<td>0.302</td>
<td>0.236</td>
<td>0.082</td>
<td>0.046</td>
<td>0.656</td>
<td>0.723</td>
<td>0.620</td>
<td>0.264</td>
<td>0.161</td>
<td></td>
</tr>
<tr>
<td>NCT Link: Random Forests</td>
<td>0.196</td>
<td>0.219</td>
<td>0.154</td>
<td>0.072</td>
<td>0.051</td>
<td>0.734</td>
<td>0.808</td>
<td>0.709</td>
<td>0.298</td>
<td>0.185</td>
<td></td>
</tr>
<tr>
<td>NCT Link: Gradient Boosting</td>
<td>0.143</td>
<td>0.162</td>
<td>0.102</td>
<td>0.046</td>
<td>0.030</td>
<td>0.717</td>
<td>0.791</td>
<td>0.684</td>
<td>0.282</td>
<td>0.182</td>
<td></td>
</tr>
<tr>
<td>+NCT Link: DHN</td>
<td>0.308</td>
<td>0.342</td>
<td>0.244</td>
<td>0.123</td>
<td>0.082</td>
<td>0.824</td>
<td>0.873</td>
<td>0.920</td>
<td>0.358</td>
<td>0.221</td>
<td></td>
</tr>
</tbody>
</table>

Results. Table 2 depicts the performance of all baseline systems as well as all configurations of NCT link when evaluated according to both judgment strategies and measured with all five metrics. As expected, all systems obtained higher performance when using the OPEN judgment scheme than when using the CLOSED scheme. The poorer performance of all systems when using the CLOSED judgment scheme supports the notion that many relevant articles were incorrectly labeled as IRRELEVANT. Consequently, the OPEN judgment scheme may be viewed as an upper bound while the CLOSED judgment scheme may be viewed as a lower bound of each system’s performance. Regardless of judgment scheme, NCT Link using the Deep Highway Network (DHN) obtains the highest performance, followed by the three other NCT Link configurations employing learning-to-rank. It is interesting to note that the complexity of these three models coincided with an increase in performance when using the OPEN judgment scheme and a decrease in performance when using the CLOSED judgment schemes. This indicates that the more complex models may have over-fit on the OPEN judgment scheme used for training. The lowest performance was exhibited by the exact match baseline, reinforcing the observations reported by Bashir et al. (2017)\(^9\) that considering the NCT ID alone is not sufficient to determine links between MEDLINE articles and clinical trials. Likewise, the disparity in performance between the basic information retrieval system (IR: BM25) and the BM25 configuration of NCT Link clearly indicates that the search criteria described in the Searching MEDLINE Articles section obtains higher quality results than when using a naïve retrieval strategy. Moreover, the increase in performance when incorporating learning-to-rank within NCT Link suggests that the features extracted by NCT Link are able to capture useful criteria for determining whether a link exists between an article and a trial. When comparing the performance of NCT Link using the DHN against the performance when using Random Forest, Linear Regression, and Gradient Boosting, it is clear that the DHN obtains superior performance, suggesting that our DHN is able to successfully extract “meta”-features capturing additional semantics about the relationship between a MEDLINE article and a clinical trial.

Discussion

We manually analyzed the MEDLINE articles retrieved by NCT Link for 30 clinical trials in test set and found four main sources of error.

The most common source of errors we observed was the result of investigator and author names. Specifically, we found that, in general, clinical trials represented investigator names with three fields: first name, middle name, and last name. However, many journals in MEDLINE only report the authors’ last names and the initials of first and sometimes middle names. This resulted in scenarios in which the system incorrectly concluded that the investigator of a trial was the same as the author of a paper. This error was most prevalent for common last names (e.g., Lin, Brown), common first initials (e.g., J, M, S, D), and when the middle initial was unspecified. Moreover, we observed that in several cases, the first and middle names in the clinical trial registry were blank and the last name contained the full name of the primary investigator. In four cases case, the first and middle names were blank and the last name appeared to refer to the sponsoring company. In future work, we believe some of these errors could be at least partially addressed by incorporating some degree of citation analysis to help (1) disambiguate initials and/or (2) infer...
unspecified names from previous work. The second most common source of errors were mismatched affiliations. We found many cases in which the same institution was referenced in multiple ways, (e.g. “UCLA” and “University of California, Los Angeles”). Moreover, addresses were often specified with different levels of detail (street names, cities, states, country). Unfortunately, resolving this kind of ambiguity is a difficult problem involving world, spatial, and geographical knowledge as well as prior knowledge about known institutions and their standard abbreviations.

The third most common source of errors was inconsistencies in the way clinical trial completion dates were provided to the registry. Because the completion date was represented in natural language, completion dates were represented in a wide variety of formats. For example, some SRMs preferred formatting the dates in the European fashion (day-month-year), while others preferred the American notation (month-day-year). In some cases, only the month and the year were indicated. Individual months were specified using digits (e.g. “01”), the full name (e.g., “January”) as well as a variety of abbreviations (e.g., “J”, “Jan”, and “Jan.”). Years were specified in both two and four digit varieties (e.g., “07”, and “2007”). In our study, we investigated applying automatic tools for recognizing time expressions (e.g., SUTime\(^{22}\)) but found it increased processing time by two orders of magnitude.

The final source of errors appears to result from SRMs providing incorrect information to the registry. We found cases in which the references provided as “result articles” for a clinical trial were published before the trial’s start date (in some cases, decades before). It is unclear whether incorrect citations were given, or whether there was confusion between the “related articles” and “result articles” fields in the registry.

In addition to the errors described above, there were some limitations in our experiments. First, we only considered the clinical trials registered on ClinicalTrials.gov despite the availability of other registries such as the World Health Organization (WHO) International Clinical Trials Registry Platform (ICTRP)*. Second, we limited our system to considering only articles published on MEDLINE and did not consider other databases such as EMBASE† or research conference proceedings. Moreover, because MEDLINE itself only provides abstracts, NCT Link did not have access to the full text of articles. In future work, it may be advantageous to consider the full text of articles included in the PubMed Central (PMC) Open Access Subset‡ (OAS); it should be noted, however, that the PMC OAS contains just over 1 million articles while MEDLINE itself contains over 14 million articles.

**Implementation Details**

NCT Link was implemented in both Java (version 8) and Python (version 2.7.13). Java was used for (1) parsing the data from MEDLINE as well as ClinicalTrials.org, relying on the Java Architecture for XML Binding (JAXB, version 2.1), (2) indexing and searching clinical trials and MEDLINE articles, relying on Apache Lucene§ (version 6.6.0), and (3) feature extraction. The Deep Highway Network (DHN) was implemented in Python using Tensorflow¶ (version 1.3). L2R baselines relied on the implementation provided by the RankLib component of the Lemur Project∥ using the recommended parameters. Our DHN used 200-dimensional internal layers. When designing the network, we found that changing the dimensionality of internal layers had no discernible effects on performance.

**Conclusion**

In this paper, we have presented NCT Link, a system for automatically linking clinical trials to MEDLINE articles reporting their results. While traditional approaches for linking trials to their publications rely on arduous, manual analyses, NCT Link learns to automatically determine the likelihood that a published article reports the results of a clinical trial by incorporating state-of-the-art deep learning and information retrieval techniques, obtaining a 30%-58% improvement over previously reported automatic systems. These promising results suggest that NCT Link will provide a useful tool for clinicians seeking to provide timely, evidence-based care. Opportunities for future work include (1) citation analyses to resolve investigator and author names, (2) geo-spatial reasoning to resolve investiga-
tor/author affiliations, (3) temporal expression normalization to account for variations in the way trial completion dates are expressed.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Predicting the Outcome of Patient-Provider Communication Sequences using Recurrent Neural Networks and Probabilistic Models

Mehedi Hasan, BS1*, Alexander Kotov, PhD1*, April Idalski Carcone, PhD2, Ming Dong, PhD1, Sylvie Naar, PhD2

1Department of Computer Science, Wayne State University, Detroit, Michigan
2Department of Family Medicine and Public Health Sciences, School of Medicine, Wayne State University, Detroit, Michigan

Abstract The problem of analyzing temporally ordered sequences of observations generated by molecular, physiological or psychological processes to make predictions about the outcome of these processes arises in many domains of clinical informatics. In this paper, we focus on predicting the outcome of patient-provider communication sequences in the context of the clinical dialog. Specifically, we consider prediction of the motivational interview success (i.e. eliciting a particular type of patient behavioral response) based on an observed sequence of coded patient-provider communication exchanges as a sequence classification problem. We propose two solutions to this problem, one that is based on Recurrent Neural Networks (RNNs) and another that is based on Markov Chain (MC) and Hidden Markov Model (HMM), and compare the accuracy of these solutions using communication sequences annotated with behavior codes from the real-life motivational interviews. Our experiments indicate that the deep learning-based approach is significantly more accurate than the approach based on probabilistic models in predicting the success of motivational interviews (0.8677 versus 0.7038 and 0.6067 F1-score by RNN, MC and HMM, respectively, when using under-sampling to correct for class imbalance, and 0.8381 versus 0.7775 and 0.7520 F1-score by RNN, MC and HMM, respectively, when using over-sampling). These results indicate that the proposed method can be used for real-time monitoring of progression of clinical interviews and more efficient identification of effective provider communication strategies, which in turn can significantly decrease the effort required to develop behavioral interventions and increase their effectiveness.

Introduction

Temporally ordered sequences of discrete or continuous observations generated by molecular, psychological or psychological process(es) arise in many different areas of biology and medicine (e.g., DNA base-pairs, protein sequences, ECG measurements, laboratory results, diagnostic codes, utterances in the clinical dialog). Classification (or categorization) is a type of analysis of those sequences that has a broad range of important practical applications, from protein function1 or structure2 prediction to detecting individuals with a heart disease3. Taking into account both the entire set of observations in a sequence, as well as the temporal order and potential dependencies between observations, makes sequence classification a more challenging task than a classification of independent observations. Predicting the outcome of those sequences (e.g. physiological or behavioral response) can also be viewed as a sequence classification problem.

In general, sequence classification methods fall into one of three major classes: feature-based, distance-based and model-based. Feature-based methods transform a sequence into a feature vector and apply a standard supervised machine learning method, such as Support Vector Machine4 or Decision Tree5 to arrive at classification decision. The methods in this class have had limited success since traditional feature representation methods cannot easily account for the order of and dependencies between observations in a sequence.

Distance-based methods classify a sequence by finding the most similar sequences with known classes based on a distance metric. The most commonly used distance metric is Euclidean distance with Dynamic Time Wrapping6. However, distance metrics are primarily designed for time series data, in which the observations are discretized by timestamps. The third type of sequence classification methods first creates a probabilistic model, such as the Markov Chain (MC) or Hidden Markov Model7 (HMM), for sequences in each class based on the training data and then, classifies new sequences by applying the created models. While MCs and HMMs can capture first- and second-order dependencies between adjacent observations in a sequence, learning higher-order dependencies with these models requires prohibitively large amounts of data. By encoding sequences into low-dimensional representations, Recurrent

* Authors provided an equal contribution.
Neural Networks (RNNs) are able to capture both short- and long-term dependencies and were shown to be effective at modeling different types of sequential data. Long Short-Term Memory (LSTM) is a variant of RNNs, which successfully addressed the vanishing gradient problem of traditional RNN. LSTM demonstrated excellent performance in different domains, from speech and handwriting recognition to health informatics. Specifically, LSTM was used as part of a multi-label classification method to recognize patterns in multivariate time series of clinical measurements, such as body temperature, heart rate and blood pressure. LSTM was also effectively used for predicting the diagnosis and medication codes, given a sequence of codes from the previous patient visits. A further simplification and improvement of LSTM model, called the Gated Recurrent Unit (GRU), was later proposed. LSTM and GRU demonstrated markedly better performance among all other RNN variants for a variety of tasks in different domains.

In this paper, we address the problem of predicting the outcome of coded patient-provider communication (PPC) sequences in the context of the clinical dialog. Specifically, we focus on predicting the success (i.e., eliciting a particular type of patient behavioral response) of motivational interviews with obese adolescents and their caregivers based on an observed sequence of coded PPC exchanges during those interviews. Childhood obesity is a serious public health concern in the United States. Recent estimates indicate that approximately one-third (31.8%) of U.S. children 2-19 years of age are overweight and 16.9% are obese. Adolescents, who are obese, are likely to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis. One approach to effective obesity intervention is Motivational Interviewing (MI), an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change. The goal of MI is to encourage patients to explore their own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements collectively referred to as “change talk” (CHT), consistently predict the actual behavior change that can be sustained for as long as 34 months after an interview. However, the ability of providers to consistently elicit this type of patient communication requires knowledge of effective communication strategies for a variety of patients, which can only be obtained through analysis of a large number of annotated interviews. Since manual examination and analysis of MI interview transcripts is a very time-consuming process, designing effective MI interventions and tailoring them to particular populations can take years. Therefore, there is a need for informatics-based methods to facilitate the development of effective behavioral interventions, in general, and theoretically-grounded computational models to explore the mechanisms of MI’s efficacy, in particular.

Our goal is to compare the accuracy of probabilistic models, such as MC and HMM, and deep learning methods, such as LSTM and GRU, for the task of predicting the success of clinical interviews (i.e., eliciting a particular type of patient behavioral response, such as CHT) at any point during a clinical interview based on a sequence of coded previous PPC exchanges in the same interview. This study is a continuation of our previous work, in which we explored several machine learning methods for automatic annotation of clinical interview fragments with a large number of patient and provider behavior codes from a specialized codebook. While there have been some previous qualitative studies of patient-provider dialog in a clinical setting, no previous work explored the applicability of state-of-the-art methods for sequence modeling to the analysis of PPC exchanges, in general, and predicting the desired patient behavioral response in the context of motivational interviews, in particular.

Methods

Data collection

The experimental dataset for this work was constructed from the transcripts of 129 motivational interviews, which consist of a total of 50,239 segmented and annotated utterances. Each transcript corresponds to an MI interview session, which typically involves a counselor, an adolescent and a caregiver. The utterances were annotated based on the MYSCOPE codebook, in which the behavior codes are grouped into the patient (adolescent and caregiver) codes and the counselor codes. Annotated utterances were divided into successful and unsuccessful communication sequences. Successful communication sequences are the ones, which resulted in positive change talk (CHT+) or commitment language (CML+) statements by an adolescent or a caregiver, while unsuccessful sequences are the ones, which resulted in negative change talk (CHT-) or commitment language (CML-), or the ones, in which no change talk or commitment language statements were made.

A fragment of an adolescent session transcript is presented in Table 1. In this example, $SS \rightarrow OQO \rightarrow HUPO \rightarrow$
OQTBN → CHT+ is a successful sequence, in which a counselor starts with an open-ended question and ultimately is able to elicit a positive change talk statement. As follows from this example, similar utterances, such as “Yeah” and “Yes”, can be assigned different behavior codes (CHT+ and HUPW), depending on the context.

Table 1: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent. MYSCOPE codes assigned to the utterances and their meaning are shown in the first two columns.

<table>
<thead>
<tr>
<th>Code</th>
<th>Behavior</th>
<th>Speaker</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>Structure Session</td>
<td>Counselor</td>
<td>Okay. Can I meet with Xxxx alone for a few minutes?</td>
</tr>
<tr>
<td>OQO</td>
<td>Open-ended question, other</td>
<td>Counselor</td>
<td>So, Xxxx, how are you doing?</td>
</tr>
<tr>
<td>HUPO</td>
<td>High uptake, other</td>
<td>Adolescent</td>
<td>Fine</td>
</tr>
<tr>
<td>OQTBN</td>
<td>Open-ended question, target behavior neutral</td>
<td>Counselor</td>
<td>That’s good. So, tell me how do you feel about your weight?</td>
</tr>
<tr>
<td>CHT+</td>
<td>Change talk positive</td>
<td>Adolescent</td>
<td>It’s not the best.</td>
</tr>
<tr>
<td>CQECHT+</td>
<td>Closed question, eliciting change talk positive</td>
<td>Counselor</td>
<td>It’s not the best?</td>
</tr>
<tr>
<td>CHT+</td>
<td>Change talk positive</td>
<td>Adolescent</td>
<td>Yeah</td>
</tr>
<tr>
<td>CQTBN</td>
<td>Closed question, target behavior neutral</td>
<td>Counselor</td>
<td>Okay, so have you tried to lose weight before?</td>
</tr>
<tr>
<td>HUPW</td>
<td>High uptake, weight</td>
<td>Adolescent</td>
<td>Yes</td>
</tr>
</tbody>
</table>

The resulting experimental dataset was highly imbalanced. Out of 5143 observed sequences, 4225 or 82.15% were positive and only 918 or 17.85% were negative. No major differences were observed in the average length of successful (9.79 utterances) and unsuccessful (9.65 utterances) sequences.

Since severely imbalanced datasets often distort the true performance of a classification method relative to a simple “majority vote” baseline (e.g. simply classifying every communication sequence as successful would result in 82.15% accuracy on our dataset), it is important to properly address the class imbalance. We evaluated the performance of probabilistic and deep learning methods using both under-sampling and over-sampling for balancing the number of samples in different classes. Synthetic Minority Over Sampling Technique (SMOTE)\textsuperscript{24} is a widely used oversampling method for imbalanced datasets, in which new synthetic examples are generated for minority classes. Specifically, we generated synthetic examples at the borderline between the majority and minority classes\textsuperscript{25}. On the other hand, the under-sampling method reduces the number of samples in majority class by replacing the clusters of samples identified by the $k$-means clustering algorithm with the cluster centroids.

**Sequence classification methods**

In general, a sequence can be viewed as a temporally ordered set of observations. In this study, an observation corresponds to a behavior code, which has a symbolic representation, such as $LUP^+$ (low uptake, positive), $OQECHT^+$ (open-ended question, elicit change talk positive), etc. Given a sequence of behavior codes $S_i = \{c_1, c_2, ..., c_n\}$ representing PPC exchanges during some part of a motivational interview, the task of predicting interview success can be considered as sequence classification. Given a set of class labels $L = \{l_1, l_2, ..., l_m\}$ (in our case, the labels are “successful” and “unsuccessful” motivational interview), a sequence classifier $C$ learns a function $S_i \rightarrow l_i, l_i \in L$ that maps a sequence $S_i$ into a class label $l_i \in L$.

Our proposed baseline prediction method consists of two steps. In the first step, we model successful and unsuccessful patient-provider interactions using first and second-order Markov Chain and Hidden Markov Model, which are popular probabilistic models for discrete observation sequences with finite vocabulary. In the second step, we classify each test sequence based on the maximum likelihood of generating that sequence from each model. Although HMM was originally developed for speech recognition\textsuperscript{7}, it is one of the most widely used methods for sequence modeling\textsuperscript{26, 27}. However, the latest advances in deep learning suggest that RNNs may provide better results than conventional machine
learning methods for the task of sequence classification. To verify this hypothesis, we employed two state-of-the-art variants of RNN in our experiments: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU).

**Markov Chain (MC)** is a probabilistic model that conditions each observation in a sequence only on preceding observation and not on any other past observation. First, we estimated two Markov models $M$ and $\overline{M}$, summarizing counselor strategies and patient responses, in the cases of successful ($M$) and unsuccessful ($\overline{M}$) motivational interviews. A Markov model $M$ can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, AMB-, LUP+, LUP-, HUPW, OQO, CQQTBN, CQECHT+, \ldots\}$ is a set of vertices, consisting of adolescent, caregiver and counselor MI behavior codes;
- $E \subseteq V \times V$ is a set of edges corresponding to possible transitions from one MI behavior code to the other in a sequence;
- $p_M : E \rightarrow [0...1]$ is a function that assigns probability $p(c_i|c_j)$ to an edge between the MI behavior codes $c_i$ and $c_j$ based on the maximum likelihood estimation:

$$P_M(c_j|c_i) = \frac{n_{c_i,c_j}}{n_{c_i}}$$  \hspace{1cm} (1)

where $n_{c_i,c_j}$ and $n_{c_i}$ are the number of times a transition between the MI behavior codes $c_i$ and $c_j$ and the number of times the code $c_i$ have been observed in the training data, respectively. Given a Markov model $M$ (such that $S \subseteq V$), the probability that a sequence of MI behavior codes $S = \{C_1, \ldots, C_N\}$ has been generated from a Markov model $M$ is:

$$P_M(S) = \prod_{i=2}^{N} p_M(c_i|c_{i-1}) = \prod_{i=2}^{N} p_M(c_i|c_{i-1})$$  \hspace{1cm} (2)

In the second step, we quantify the likelihood of success of a given motivational interview at a certain time point given a sequence of MI behavior codes $S$ observed prior to that point as:

$$p(S \rightarrow success\ ful) = \log \left( \frac{P_M(S)}{P_{\overline{M}}(S)} \right) = \sum_{i=2}^{N} \log p_M(c_i|c_{i-1}) - \sum_{i=2}^{N} \log p_{\overline{M}}(c_i|c_{i-1})$$  \hspace{1cm} (3)

If $p(S \rightarrow success\ ful) > 0$, a communication sequence is predicted to be successful (i.e. result in positive change talk or commitment language). Otherwise, it is predicted to be unsuccessful.

The above model is also referred as first-order MC, since it only considers immediately preceding behavior code, when computing the state transition probabilities. In our experiment, we also considered second-order Markov model, which conditions each observation on the preceding two observations.

**Hidden Markov Model (HMM)** is another probabilistic model used for modeling processes varying in time. HMMs are widely used for sequence analysis because of their ability to identify hidden states, corresponding to clusters of observations. Mathematically, HMM can be defined as $\lambda = (A, B, \pi)$, where:

- $A$ is an $N \times N$ state transition probability distribution matrix $A = \{a_{ij}\}$
- $B$ is an $N \times M$ matrix $B = \{b_j(k)\}$ with observation symbol probability distribution for each state
- $\pi$ is the initial state distribution vector $\pi = \{\pi_i\}$

Hence, $N$ is a number of hidden states in the model and $M$ is a number of distinct observations per hidden state, i.e. the discrete vocabulary size. The key difference between HMM and MC is that HMM requires specifying the number of hidden states as a model parameter. HMM deduces a sequence of hidden states that best explains the observations along with the state transition probabilities and the distributions of observations (emission probabilities) per each hidden state. The Baum-Welch algorithm is used to estimate the parameters of HMMs for successful and
unsuccessful interviews using the corresponding training set, while the Viterbi algorithm\textsuperscript{7} is used to determine the most likely sequence of hidden states for a given sequence of observations. After assignment of hidden states, the log-likelihood of success for an interview can be estimated using Eq. 3 as well.

**Behavior code embeddings.** Representation of behavior codes was inspired by the recent success of word embeddings\textsuperscript{28–30}. Embedding is a representation of an object in low-dimensional space using a real-valued vector. In our study, embeddings of behavior codes were obtained as a by-product of training LSTM and GRU after feeding one-hot vectors as a representation of behavior codes as input to these RNNs. Behavior code embeddings have the property of representing similar codes with the vectors that are close to each other in low-dimensional space. Figure 1 illustrates the MYSCOPE code embeddings visualized in 2-dimensional space by t-SNE\textsuperscript{31}. It can be seen that positive behavior codes such as OQECHT+, OQECML+, AF, AFL, SUP, RCML+S, CQECML+, etc. formed a cluster in the left part of Figure 1. The nearest neighbors of CQECML+ are highlighted by different color intensity (i.e. OQECML+ being more purple indicates that it is more similar to CQECML+). The right part of the figure demonstrates another cluster formed with negative behavior codes including CQECML-, AMB-, RCHT-C, OQECHT-, GINFO-, RBAC, LUP-, RCHT-S, RPTBC, RAMBC, AMB-, RCML-S, etc. It is interesting that the behaviors intended to elicit CHT+/CML+ group together, whereas the ones intended to elicit CHT-/CML- also group together and are located in the opposite regions of semantic space.

Recurrent Neural Networks (RNN) are a class of neural networks that have an internal memory, which makes them particularly suitable for processing sequences of observations. The ability of RNNs to capture long-term dependencies and remember past observations for predicting future observations is their main advantage over MCs and HMMs. These features are very useful in the analysis of motivational interviews, in which any behavior observed at a particular point in the interview may be indicative of other behaviors that are observed later. In order to mitigate the vanishing gradient problem of earlier versions of RNN\textsuperscript{10}, Hochreiter et al.\textsuperscript{9} proposed Long Short Term Memory networks (LSTM). There are several variants of LSTM model, among which the most notable one is the Gated Recurrent Unit\textsuperscript{32} (GRU). GRUs are simpler than LSTMs and have been shown to be effective for a variety of Natural Language Processing tasks\textsuperscript{32}. GRU is formally defined as follows:

\[
z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{4}
\]
\[
r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{5}
\]
\[
\tilde{h}_t = \tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{6}
\]
In Eq. 4-7, $\sigma$ corresponds to sigmoid function and $\odot$ designates an element-wise product. The update gate $z_t$ and reset gate $r_t$ at time step $t$ are computed by the Eq. (4) and (5), where $W_z, W_r, W_h, U_z, U_r, U_h$ are the weight matrices and $b_z, b_h, b_r$ are bias vectors. The activation $h_t$ of the GRU at time $t$ is a linear combination of the previous activation $h_{t-1}$ and the candidate activation $\tilde{h}_t$, which is represented by Eq. (7) and (6).

\[
h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t
\]

(7)

Figure 2: Proposed RNN model with target replication (TR).

The RNN architecture employed for sequence classification is shown in Figure 2. As can be seen from Figure 2, softmax is used at each time step to predict the class of a sequence observed so far. Since the sequence label is predicted at each observation, the proposed architecture is referred to as Recurrent Neural Network with Target Replication (TR). It is trained by minimizing the following hybrid loss function:

\[
\tilde{L} = \alpha \cdot \frac{1}{T} \sum_{t=1}^{T} L(\tilde{y}^{(t)}, y^{(t)}) + (1 - \alpha) \cdot L(\tilde{y}^{(T)}, y^{(T)})
\]

(8)

As follows from Eq. 8, the total loss $\tilde{L}$ is a convex combination of the final loss $L(\tilde{y}^{(T)}, y^{(T)})$ and the average loss over all observations in a sequence, where $T$ is the total number of observations, $\tilde{y}^{(t)}$ is the output at step $t$, and $\alpha \in [0, 1]$ is a hyperparameter controlling the relative importance of each loss type. We experimentally determined that the best performance is achieved when $\alpha = 0.5$. Our model also contains several other hyperparameters, such as the number of embedding dimensions, the number of hidden units, learning rate, batch size, etc., which were optimized on the validation set. We implemented our models in Tensorflow with Adam optimizer as well as early stopping based on the validation loss and observed that our model converges after 100 epochs.

Evaluation metrics

Performance of probabilistic and deep learning methods* was evaluated in terms of precision, recall, and F-measure using 10 folds cross-validation and weighted macro-averaging of these metrics over the folds. However, LSTM and GRU were trained on 80% of the data and validated on 10%, with the remaining 10% of the data used for testing.

Results

All sequence classification methods were evaluated in the case of both under and over-sampling. Predictive performance summary of all methods is summarized in Table 2.

*the code for all models is publicly available at https://github.com/teanalab/myscope-sequential-analysis

69
Table 2: Performance of MC, HMM, LSTM and GRU with and without target replication (TR) for predicting the success of patient-provider communication sequences when under- and over-sampling were used to balance the dataset. The highest value for each performance metric is highlighted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Under-sampling</th>
<th>Over-sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Markov Chain 1st Order</td>
<td>0.7060</td>
<td>0.7044</td>
</tr>
<tr>
<td>Markov Chain 2nd Order</td>
<td>0.6395</td>
<td>0.6385</td>
</tr>
<tr>
<td>Hidden Markov Model</td>
<td>0.6244</td>
<td>0.6143</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.8672</td>
<td>0.8626</td>
</tr>
<tr>
<td>LSTM-TR</td>
<td><strong>0.8733</strong></td>
<td><strong>0.8681</strong></td>
</tr>
<tr>
<td>GRU</td>
<td>0.8674</td>
<td>0.8648</td>
</tr>
<tr>
<td>GRU-TR</td>
<td>0.8705</td>
<td>0.8676</td>
</tr>
</tbody>
</table>

**Predictive performance in the case of under-sampling**

We used a small learning rate of 0.00005 and the batch size of 8 along with early stopping strategy for training deep learning models on the dataset balanced with under-sampling. Five major conclusions can be drawn from the results in Table 2. First, recurrent neural networks outperform probabilistic models and achieve 16.39%-26.1% higher F1-score. Second, LSTM with target replication has the best performance over all other RNN methods, and achieved F1-score 0.8677 with precision 0.8733 and recall 0.8681. Third, target replication strategy improves the performance of GRU and LSTM, with conventional GRU showing better performance than traditional LSTM. Fourth, among probabilistic models, the MC based method generally outperforms HMM across all metrics for under-sampled sequences. Fifth, second-order MC has lower precision, recall, and F-measure than first-order MC. In particular, precision, recall and F-measure decrease by 9.42%, 9.36% and 9.36%, when going from first to second-order MC model.

**Predictive performance in the case of over-sampling**

Similar to the under-sampling scenario, early stopping strategy was also employed for training deep learning models on the dataset balanced with over-sampling. However, in this case, RNN models were trained with the learning rate of 0.00010 and the batch size of 55. Experimental results indicate that HMM had better performance than second-order MC, achieving 9.34%, 7.65%, and 7.43% higher precision, recall, and F-measure, while HMM still had 1.98%, 2.97%, and 3.28% lower precision, recall, and F-measure than first-order MC. Also similar to the under-sampling scenario, target replication improves the performance of RNN models and LSTM with target replication has the highest F1-score among all models. However, the predictive performance of LSTM and RNN decreases when over-sampling is used, while the performance of probabilistic models increases.

Table 3: Most likely communication sequences in successful and unsuccessful motivational interviews.

<table>
<thead>
<tr>
<th>Type</th>
<th>Most likely communication sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>successful</td>
<td>GINFO+: General information, positive → LUP+: Low uptake, positive → OQTBN: Open-ended question, target behavior neutral</td>
</tr>
<tr>
<td>successful</td>
<td>SS: Structure session → GINFO+: General information, positive → CQECHT+: Closed-ended question, elicit change talk positive</td>
</tr>
<tr>
<td>successful</td>
<td>SO: Statement, other → LUP+: Low uptake, positive → AF: Affirm → HUPW: High uptake, weight → OQECLM+: Open-ended question, elicit commitment language positive</td>
</tr>
<tr>
<td>unsuccessful</td>
<td>ADV+: Advise, positive → AMB-: Ambivalence negative → OQECHT+: Open-ended question, elicit change talk negative</td>
</tr>
<tr>
<td>unsuccessful</td>
<td>CQECHT+: Open-ended question, elicit change talk positive → RCHT-S: Reflect, change talk negative → OQECHT+: Open-ended question, elicit change talk negative</td>
</tr>
<tr>
<td>unsuccessful</td>
<td>SUP: Support → AF: Affirm → CQTBN: Closed-ended question, target behavior neutral → OQECHT+: Open-ended question, elicit change talk negative → AMB-: Ambivalence negative</td>
</tr>
</tbody>
</table>
Most likely communication sequences

Table 3 provides examples of typical patient-provider communication sequences that frequently appear in successful and unsuccessful motivational interviews. We observed that in successful motivational interviews information is frequently provided using patient-centered communication (GINFO+) and structure session (SS) utterances, in which the counselor either explains the therapeutic agenda or attempts to transition to a new topic or session content. Sometimes, counselors also acknowledge the clients’ communication or an off topic comment (SO). We also observed that affirmations (AF) and open-ended questions (OQECML+) have a strong effect on eliciting positive change talk or commitment language, which is consistent with MI theory. It can also be seen that providing advice using non-patient centered strategies (ADV-) leads to negative ambivalence (AMB-), which results in the interview heading in therapeutically wrong direction. Questions posed to elicit negative change talk or commitment language lead to CHT-, CML- or AMB-, which is consistent with the manual analysis by clinicians.

Discussion

By analyzing the experimental results of different communication sequence outcome prediction methods proposed in this paper, we arrived at the following conclusions. First, the overall predictive performance of RNN based methods is substantially higher than that of probabilistic models. In particular, the RNN-based methods achieve near-human accuracy for predicting the success of motivational interviews. This indicates that RNN is able to capture the structure of discourse in motivational interviews by preserving long-term dependencies among the behavior codes, which reflect the overall progression of the interviews. This provides evidence that RNNs are able to successfully replicate human cognitive processes to integrate previous information when making decisions. In addition to that, embeddings allow to reduce the dimensionality of codes in PPC sequences and consequently improve both precision and recall of prediction. Second, using target replication to compute the loss at each time step results in better performance for all configurations of the proposed RNN-based methods. This indicates that the average of the losses over all steps emphasizes the dependencies between the pairs of patient and provider codes, which results in more accurate estimates of the model parameters. Better estimates of parameters in RNN models of motivational interviews are propagated to the next step based on the relative importance of intermediate output, where they are aggregated into predictions for the entire sequence. This allows to achieve an improvement in prediction accuracy.

Third, using first-order Markov model results in better prediction accuracy compared to higher-order Markov models, which we attribute to the fact that the number of states in higher-order Markov models may grow exponentially with their order. As a result, accurate estimation of transition probabilities requires much larger training data. Using smaller datasets, which is the case when under-sampling is employed, will result in a sparsity problem, when many transitions are either not observed in the training set at all or observed only a few times, leading to missing or potentially inaccurate probability estimates. Obtaining large training sets cannot be easily accomplished in many domains, including motivational interviewing. In this study, we found out that using first-order Markov models is a reasonable trade-off between efficiency and accuracy.

Fourth, similar to traditional Markov model, HMM achieves a dramatic improvement in the prediction accuracy when larger training set is used. This indicates that sufficient training data is required to find the optimal settings of hyperparameters, such as the number of hidden states, initial state distribution, transition probabilities, and emission probabilities.

Fifth, the proposed method can be used to identify the most effective communication strategies for eliciting a particular type of behavioral response. Awareness of these strategies by researchers can significantly decrease the time and effort required to develop effective interventions to address many public health conditions, such as childhood obesity, and tailor these interventions to particular patient cohorts. Awareness of these strategies by the counselors can lead to a greater success rate of motivational interviews.

Conclusion

In this paper, we compared the accuracy of Recurrent Neural Networks with Markov Chain and Hidden Markov Model for the task of predicting the success of motivational interviews. We found out that individual PPC exchanges
are highly indicative of the overall progression and future trajectory of clinical interviews and can be used to predict their overall success. Our proposed methods can facilitate motivational interviewing researchers in establishing causal relationships between different communication strategies and the desired behavioral outcomes during the interviews without resource-intensive manual qualitative analysis of interview transcripts, which can significantly decrease the time and effort required to develop behavioral interventions. Our proposed methods can also help to identify the most likely sequences in successful and unsuccessful motivational interviews, which can directly inform clinical practice and increase the effectiveness of behavioral interventions. Our experimental results also indicate that the proposed methods can be used for real-time monitoring of the progression of clinical interviews. This work also has broad implications for public health research by providing a theoretically-grounded computational approach to qualitative data analysis.

Acknowledgments
This study was supported by a grant from the National Institutes of Health, NIDDK R21DK108071, Carcone and Kotov, MPIs. We would like to thank the student assistants in the Department of Family Medicine and Public Health Sciences at Wayne State University School of Medicine for their manual annotation of experimental dataset with MYSCOPE codes.

References


Integrating Structured and Unstructured EHR Data Using an FHIR-based Type System: A Case Study with Medication Data

Na Hong, Andrew Wen, Feichen Shen, Sunghwan Sohn, Sijia Liu, Hongfang Liu, Guoqian Jiang
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract
Standards-based modeling of electronic health records (EHR) data holds great significance for data interoperability and large-scale usage. Integration of unstructured data into a standard data model, however, poses unique challenges partially due to heterogeneous type systems used in existing clinical NLP systems. We introduce a scalable and standards-based framework for integrating structured and unstructured EHR data leveraging the HL7 Fast Healthcare Interoperability Resources (FHIR) specification. We implemented a clinical NLP pipeline enhanced with an FHIR-based type system and performed a case study using medication data from Mayo Clinic’s EHR. Two UIMA-based NLP tools known as MedXN and MedTime were integrated in the pipeline to extract FHIR MedicationStatement resources and related attributes from unstructured medication lists. We developed a rule-based approach for assigning the NLP output types to the FHIR elements represented in the type system, whereas we investigated the FHIR elements belonging to the source of the structured EMR data. We used the FHIR resource “MedicationStatement” as an example to illustrate our integration framework and methods. For evaluation, we manually annotated FHIR elements in 166 medication statements from 14 clinical notes generated by Mayo Clinic in the course of patient care, and used standard performance measures (precision, recall and f-measure). The F-scores achieved ranged from 0.73 to 0.99 for the various FHIR element representations. The results demonstrated that our framework based on the FHIR type system is feasible for normalizing and integrating both structured and unstructured EHR data.

Introduction
With the widespread usage of electronic health records (EHRs) in healthcare organizations, there are huge requirements for semantic interoperability and computable phenotyping in clinical and translational research. The lack of EHR data interoperability between institutions, however, makes it challenging for secondary use of EHR data, especially in collaborative research across institutions. Representing EHR data using a standard data model would assist in achieving large-scale data research collaboration, and meanwhile support the rapid generation of computable phenotypes.

Several data models have been developed to provide a standardized data representation for EHR data, including, amongst others, the HL7 Consolidated Clinical Document Architecture (CDA), the HL7 Reference Information Model (RIM), the OHDSI Common Data Model (CDM), the National Quality Forum (NQF) Quality Data Model (QDM), the Informatics for Integrating Biology and the Bedside (i2b2), the HL7 Fast Healthcare Interoperability Resources (FHIR). Meanwhile, clinical Natural Language Processing (NLP) plays an important role in extracting information from clinical text to a structured representation. A number of standards-based data modeling approaches have been studied using a variety of NLP technologies, such as using NLP to automatically generate entry level CDA, mapping textual queries to OHDSI CDM, using NLP for clinical research on i2b2 data sets, using NLP supported diagnostic criteria QDM representation. Some research has discussed using NLP with output to the FHIR data model, e.g. integration of NLP data with other sources of cancer data to build FHIR-based cancer phenotypes. These studies, however, are limited in application to a specific clinical domain. To the best of our knowledge, a generic integration approach for modeling EHR data with the FHIR data model has not yet been well studied.

As an emerging next generation standard framework, FHIR was developed to meet clinical interoperability needs, and has the benefit of being relatively easy to implement and rapid to deploy. In addition, FHIR leverages the latest web standards and places a strong focus on implementability. Meanwhile, in regards to the NLP modules and tools developed independently in the NLP research community, there are broad requirements for achieving direct interoperability through a common type system with integrating applications. To extract structured data from unstructured data so as to model EHR data into the FHIR data model, a computationally processable type system that is shareable amongst multiple NLP systems is required. The Apache-licensed project Unstructured Information
Management Architecture (UIMA) provides such a software framework for building type systems while supporting interaction between multiple NLP components.

Another significant challenge is related to the terminology binding. Many elements in the FHIR resources have a coded value; some in the form of a fixed string (a sequence of characters) assigned as one of a set of fixed values defined in the FHIR specification; some in the form defined as “concept” where external terminologies or ontologies (e.g. LOINC, RxNorm or SNOMED CT) are used. In some cases, a locally maintained dictionary and/or look up table are even used as a part of FHIR profiles. Normalizing non-standard data into the coded FHIR fields thus poses a challenge.

Therefore, the overall objective of our study is to develop and evaluate a generic framework and accompanying methods for integrating structured and unstructured EMR data to a standard and interoperable format. We implement a clinical NLP pipeline enhanced with an FHIR-based type system and perform a case study using medication data from Mayo Clinic’s EHR.

Materials and Methods

Materials

Data corpus

We used a data set that was composed of 14 clinical notes randomly selected from those generated by Mayo Clinic and contained 166 manually annotated medication statement resources and their associated elements. The annotation set was used to evaluate the performance of our methods.

UIMA and UIMA-based NLP tools

UIMA is a data-driven architecture where individual components are able to communicate with one another through a data structure called the Common Analysis System (CAS), which uses a specified hierarchical type system. A common type system was thus defined under UIMA to meet the need for interoperability between different NLP12. As there are many clinical NLP tools developed on UIMA architecture, our NLP implementation was also designed using UIMA so as to increase interoperability. Currently, UIMA based clinical NLP tools include: 1) cTAKES13- cTAKES provides the pipeline for selecting which descriptors are used together to support the clinical NLP tasks; 2) MedXN14- MedXN is a medication entity and attribute extraction and normalization NLP tool; and 3) MedTime15- MedTime extracts and normalizes TIMEX3-based temporal expressions from clinical text.

FHIR Specification

FHIR is based on a notion of “resource.” The FHIR specification defines a set of core resources and an infrastructure for handling resources16. The core FHIR resources represent a wide range of healthcare related concepts, both clinical and administrative. Through aggregating granular clinical concepts, the resources can support the representation of complex clinical scenarios. Currently, FHIR has 3 available exchange formats: JSON, XML and RDF17.

HAPI FHIR

HAPI FHIR is an open-source implementation of the FHIR specification in Java18. HAPI FHIR defines model classes for every resource type and data type defined by the FHIR specification. The class FhirContext acts as a factory for most other parts of the API as well as a runtime cache of information that HAPI needs to operate. As the HAPI FHIR APIs supports the data model as defined by FHIR specifications, we used the HAPI FHIR API to serialize data stored in our UIMA-based FHIR type system into standard FHIR XML and JSON representations.

Methods

For each different types of FHIR resource, corresponding EHR data may exist as either structured data, e.g. patient demographics and laboratory test observations stored within a database, or as unstructured clinical notes, e.g. medication lists and problem lists. In many cases, the unstructured data is embedded within semi-structured EHR data, which provides some information on the nature of the data’s content. According to different characteristics of the EHR data, we designed an integrated framework for modeling EHR data into the FHIR specification. The FHIR data modeling tasks are separated into two workflows, as shown in Figure 1. The first workflow is for modeling...
unstructured EHR data in FHIR. This requires using different clinical NLP tools to recognize elements from different clinical note sections, and to transform these clinical elements into their respective elements within the FHIR data model. The second workflow combines information from structured data with the NLP output to build a complete FHIR resource. For this workflow, the metadata mapping and data standardization are the core tasks before conducting data transformation. An FHIR-based type system is developed to integrate different FHIR elements from unstructured structured data into a generic framework that supports direct generation of the data in FHIR standard format.

![Diagram of FHIR framework](image)

**Figure 1.** The framework for integrating structured and unstructured EHR data using FHIR

Considering medication statement data is often recorded within clinical narratives, in this paper, we discuss our methods in detail using one type of FHIR resource, “MedicationStatement”. To demonstrate our expected modeling results, a standard FHIR representation of the textual medication statement “Guaifenesin 100-mg/5 mL give 10 mL every six hours as-needed for cough.” is shown in Figure 2.

```
{
  "resourceType": "MedicationStatement",
  "meta": {
    "versionId": "1",
    "lastUpdated": "2023-01-01"
  },
  "status": "draft",
  "code": {
    "coding": [
      {
        "code": "53932"
      }
    ],
    "text": "Guaifenesin"
  },
  "subject": {
    "reference": "Patient/123456"
  },
  "dosage": [1]
}
```

**Figure 2.** An example of an FHIR MedicationStatement instance
Transforming FHIR resources into annotation schemas

An annotation schema is necessary for annotating unstructured data elements, so as to guide the extraction of entities from textual notes. The FHIR RDF specification was loaded as a schema for annotation purpose, and a Protégé plugin annotation tool Knowtator\textsuperscript{19} was used to annotate the textual clinical notes. The annotation process was to identify and categorize the annotated fields into the FHIR-based annotation schema. To focus the annotation field and create operational annotation guideline, we tailored the FHIR annotation schema for sectional annotation tasks. The clinically relevant sections defined in HL7 CDA\textsuperscript{20} are used for identifying target clinical notes sections. For example, the sections contain Allergies, Medications, Problems, immunizations, Diagnostics, etc., in this study, we used the FHIR “MedicationStatement” resource and associated resources for the annotation tasks. In a word, instances of FHIR elements belonging to the FHIR MedicationStatement were manually annotated for evaluation.

Acquiring FHIR elements from unstructured data

A number of different clinical NLP tools were integrated into the overall pipeline to perform different specialized tasks depending on the specific clinical note section. In our use case of extraction information from a medication list, we recognized 3 subtasks for which we integrated 3 different NLP tools (Figure 3). MedXN was responsible for the extraction of standardized medication concept mentions as well as its related attributes; MedTime was used to extract temporal elements defined in FHIR; and finally, other entities that could not be directly extracted by any existing NLP tools, such as Dosage.additionallInstruction and MedicationStatement.reasonCode, were directly extracted from free-text through the development of specific NLP entity extraction modules.

![Figure 3](image-url)

**Figure 3.** Populating the FHIR resource “MedicationStatement” from a textual medication list

**Mapping the NLP output type to FHIR element type**

As the output type systems of the different NLP tools all differed from our target FHIR element definition, we created a set of rules to classify the detected NLP elements. In terms of the mapping analysis of FHIR elements from the output of MedXN and MedTime, we concluded that while some of the NLP output types could be directly mapped to FHIR elements without semantic differences, e.g. MedXN:Medication.form maps directly to
deemed sufficiently consistent (N.H., then used section system. We used W SNOMED CT and other
normalization work has been "by mouth", (MedicationStatement.route element
Standardization existing completed | entered (Patient|Group)
attri
instance data
Set
and imported from EHR structured tables.
Medication.form, Dosage.asneededBoolean, Medicat
an example her
the majority of the
In regards to the
Combini
present
corresponding
corresponding
corresponding
MedKAT,
UIMA
be
rapid
with the exception of certain UIMA reserved words such as begin, end, and start, for which
corresponding FHIR fields were renamed from “{field}” to “fhir{field}”), structure hierarchy, and data restrictions
present within the FHIR specification.
Combining structured data with NLP output
In regards to the unstructured notes within the EHR, mentions extracted from the text using NLP tools comprised the
majority of the content for the corresponding FHIR resource backbone. There were still, however, several pieces of
information that needed to be captured from structured EHR data and integrated with the NLP output to complete
the population of the corresponding FHIR resource’s content. We take the FHIR resource “MedicationStatement” as
an example here. While certain elements could be mapped from NLP output types, e.g. MedicationStatement.medicationCodableConcept,
Medication.form, Dosage.asneededBoolean, some of the other FHIR elements, such as MedicationStatement.status,
MedicationStatement.subject, MedicationStatement.taken, MedicationStatement.category, could be directly mapped
and imported from EHR structured tables. The key steps for combining structured data with NLP output were to: 1) Set
templates for mapping the database metadata to the corresponding FHIR resource elements 2) extract the
instance data; 3) Link the structured instance data with NLP output through a primary key reference or directly as an
attributes defined within the FHIR resource. When populating each FHIR “MedicationStatement” resource instance,
for example, we could directly get its “subject” (Who is/was taking the medication) information from structured
EHR, and linked each “subject” to the specific “MedicationStatement” instance through the Reference
(Patient|Group) identifier of the FHIR resource Patient or the FHIR resource Group. “Status” information (active |
completed | entered-in-error | intended | stopped | on-hold), as another example, could be extracted directly from
existing EHR tables but needed to be normalized using the FHIR defined value set.
Standardization and FHIR representation
FHIR names its methods of defining codes collectively as "code systems". For example, for the MedicationStatement.route element, it specifies how a drug enters the body, the “SNOMED CT Route Codes” (http://hl7.org/fhir/ValueSet/route-codes) are used by FHIR to standardize different textual mention. For example, “by mouth”, “sprays each nostril”, and “topically” would be normalized to “Oral route|26643006”, “Nasal route|46713006”, and “topical route|6064005” with their associated SNOMED CT codes. A part of the normalization work has been achieved by NLP tools: for instance, MedXN converts medication information to the RxNorm standard and maps it to the corresponding RxCUI; cTAKES extracts medical concepts and assigns UMLS concept unique identifiers (CUIs) and SNOMED CT codes by integrating dictionary lookups for the UMLS, SNOMED CT and other dictionaries as part of its name entity recognition pipeline.
Evaluation design
We evaluated the overall performance of our rule-based classification algorithm for populating the FHIR type
system. We used individual MedicationStatement resource instances ($n = 166$) from a collection of medication list
sections from 14 clinical notes. These manually annotated MedicationStatements and their related attributes were
then used as the gold standard to evaluate the performance of the rule based classification algorithm. Two authors
(N.H., F.S.) annotated these medication lists, and the inter-observer agreement achieved was 0.95, which was
deemed sufficiently consistent. Three standard measures were used to measure the performance of the extractors:
precision (P), recall (R), and F-measure (F), where \( P = \frac{TP}{TP+FP}, \) \( R = \frac{TP}{TP+FN}, \) and \( F = \frac{2PR}{P+R}, \) where TP stands for True Positive, FP stands for False Positive, and FN stands for False Negative. For the capture of structured data, a review on the element definitions of the FHIR resource was used to evaluate the feasibility of gathering the FHIR elements from the structured Mayo Clinic EHR data.

**Results**

As an initial implementation, our annotation task focused on building the FHIR resource “MedicationStatement” and its related resources and elements. A screenshot of the FHIR annotation schema in Protégé for a MedicationStatement annotation is shown in Figure 4.

![FHIR Annotation Schema](image.png)

**Figure 4.** The FHIR annotation schema in Protégé (MedicationStatement)

Table 1 shows our mapping relations between the MedXN and MedTime types and the FHIR elements in the resource “MedicationStatement”. After the model analysis, we found that the data elements mapping relations were generally 1:n maps with a few 1:1 maps, due to the FHIR model describing resources on a more refined granularity. We use abbreviation for the multi-level FHIR name system for display in Table 1. For example, the FHIR element “Dosage.timing.repeat.frequency” is displayed as “Dosage.frequency”.
Table 1. Model Elements Mapping Between MedXN/MedTime and FHIR

<table>
<thead>
<tr>
<th>NLP tools and Output Types</th>
<th>FHIR Elements (MedicationStatement)</th>
<th>Mapping Rules and Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedXN:Drug</td>
<td>MedicationStatement.medication CodableConcept</td>
<td>1:1 Example: Guaifenesin [RxCUI : 5032]</td>
</tr>
<tr>
<td>MedXN: Drug:attributes:type=&quot;frequency&quot;</td>
<td>Dosage.frequency Dosage.frequencyMax Dosage.period Dosage.periodMax Dosage.periodUnit Dosage.asNeededBoolean Dosage.dayOfWeek Dosage.when</td>
<td>1:n Examples: once daily → 1[frequency], 1[period], d[periodUnit] 4-6 times → 4[frequency], 6[frequencyMax] Regular: once daily every six hours Irregular: as needed for pain</td>
</tr>
<tr>
<td>MedXN: Drug:attributes:type=&quot;duration&quot;</td>
<td>Dosage.duration Dosage.durationMax Dosage.durationUnit</td>
<td>1:n Example: 3 days → 3[duration], d[durationUnit]</td>
</tr>
<tr>
<td>MedXN: Drug:attributes:type=&quot;strength&quot;</td>
<td>Medication.ingredient.amount. numerator.quantity.value Medication.ingredient.amount. numerator.quantity.unit Medication.ingredient.amount. denominator.quantity.value Medication.ingredient.amount. denominator.quantity.unit</td>
<td>1:n Examples: Regular: 500 mg /5 ml → 500[numerator.quantity.value], mg[numerator.quantity.unit], 5[denominator.quantity.value], ml [denominator.quantity.unit] Irregular: 200 mg → Default assign: 1[denominator.quantity]</td>
</tr>
<tr>
<td>MedXN: Drug:attributes:type=&quot;dosage&quot;</td>
<td>Dosage.doseQuantity.value Dosage.doseQuantity.unit Dosage.doseQuantity.Range.low.value Dosage.doseQuantity.Range.low.unit Dosage.doseQuantity.Range.high.value Dosage.doseQuantity.Range.high.unit</td>
<td>1:n Examples: 10 ml → 10[value], ml[unit] 2-3 tabs → 2[range.low.value], tab[range.low.unit], 3[range.high.value], tab[range.high.unit] Notes: we currently both map “tab” to dose unit and medication form</td>
</tr>
<tr>
<td>MedTime: MedTimex3:type=&quot;DATE&quot;</td>
<td>MedicationStatement.effectiveDateTime</td>
<td>1:1 Examples: April 16th</td>
</tr>
</tbody>
</table>
The FHIR Type system in XML format used for interoperability with existing NLP tools was built on the FHIR STU 3 v1.8.0 specification, which includes the full structure definition of FHIR specification, a fragment is displayed in Figure 5.

```xml
<typeDescription>
  <name>org.hl7.fhir.MedicationStatement</name>
  <description/>
  <supertypeName>org.hl7.fhir.DomainResource</supertypeName>
  <features>
    <featureDescription>
      <name>identifier</name>
      <description/>
      <rangeTypeName>uuid:5:8Array</rangeTypeName>
      <elementType>org.hl7.fhir.Identifier</elementType>
    </featureDescription>
    <featureDescription>
      <name>status</name>
      <description/>
      <rangeTypeName>org.hl7.fhir.MedicationStatementStatus</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>medicationCodeableConcept</name>
      <description/>
      <rangeTypeName>org.hl7.fhir.CodeableConcept</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>medicationReference</name>
      <description/>
      <rangeTypeName>org.hl7.fhir.Reference</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>subject</name>
      <description/>
      <rangeTypeName>org.hl7.fhir.Reference</rangeTypeName>
    </featureDescription>
    <featureDescription>
      <name>effectiveDateTime</name>
      <description/>
    </featureDescription>
  </features>
</typeDescription>
```

**Figure 5.** The FHIR type system in UIMA

We used RxNorm to normalize medication concepts, and used the FHIR-defined value sets to normalize FHIR attribute elements. Currently, we have normalized 11 FHIR elements using locally configured data constraints, with work in progress on all FHIR elements standardization work, dependent to a large extent on the FHIR terminology services interfaces. Currently normalized elements comprise:
- `MedicationStatement.medicationCodeableConcept(RxNorm)`,
- `Dosage.timing.frequency(integer)`,
- `Dosage.timing.repeat.duration (decimal)`
- `Dosage.timing.repeat.durationUnit (code: s | min | h | d | wk | mo | a - unit of time (UCUM))`
- `Dosage.timing.repeat.frequencyMax(integer)`,
- `Dosage.timing.repeat.period(decimal)`,
- `Dosage.timing.repeat.periodMax(decimal)`,
- `Dosage.timing.repeat.periodUnit(code: s | min | h | d | wk | mo | a - unit of time (UCUM))`
- `Dosage.asNeededBoolean(boolean)`,
- `Dosage.timing.repeat.dayOfWeek(code: mon | tue | wed | thu | fri | sat | sun)`

**Evaluation Results**

In our experiment, we tested the results of integrating the MedXN and MedTime NLP pipelines to populate the FHIR resource “MedicationStatement”. The standardization processes were also evaluated as part of this process. Through a comparison of the results of using rule-based classification and our manual annotation (which included the normalized form, if applicable) of each FHIR element within the medication list notes of Mayo Clinic, we evaluated the performance of our NLP pipeline. Evaluation results are shown in Table 2. The evaluation results show that our rules-based method has slightly higher precision, recall, and F-measure comparing with MedXN and MedTime for those 1:1 mapping elements. For example, the element `MedicationStatement.medicationCodeableConcept` is the case and the reason is probably because the size of our annotated data set is relatively small. However, for those 1:n mapping elements (e.g.,
Medication.ingredient.amount.quantity.value/unit (MedXN:strength)), the overall precision, recall, and F-measure is slightly lower than others, and the reason is because the entity normalization processing may cause the transforming mistakes to some extent.

<table>
<thead>
<tr>
<th>FHIR Elements</th>
<th>Total</th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedicationStatement.medicationCodeableConcept</td>
<td>166</td>
<td>0.996</td>
<td>0.982</td>
<td>0.988</td>
</tr>
<tr>
<td>Dosage.timing.repeat.frequency</td>
<td>87</td>
<td>0.795</td>
<td>0.873</td>
<td>0.832</td>
</tr>
<tr>
<td>Dosage.timing.repeat.period</td>
<td>133</td>
<td>0.959</td>
<td>0.914</td>
<td>0.936</td>
</tr>
<tr>
<td>Dosage.timing.repeat.periodUnit</td>
<td>129</td>
<td>0.951</td>
<td>0.943</td>
<td>0.947</td>
</tr>
<tr>
<td>Dosage.timing.repeat.duration</td>
<td>5</td>
<td>0.600</td>
<td>1</td>
<td>0.750</td>
</tr>
<tr>
<td>Dosage.route</td>
<td>143</td>
<td>0.957</td>
<td>0.816</td>
<td>0.878</td>
</tr>
<tr>
<td>Medication.ingredient.amount.numerator.quantity.value</td>
<td>155</td>
<td>0.930</td>
<td>0.815</td>
<td>0.869</td>
</tr>
<tr>
<td>Medication.ingredient.amount.numerator.quantity.unit</td>
<td>135</td>
<td>0.926</td>
<td>0.899</td>
<td>0.911</td>
</tr>
<tr>
<td>Medication.form</td>
<td>127</td>
<td>0.871</td>
<td>0.704</td>
<td>0.779</td>
</tr>
<tr>
<td>Dosage.dose.quantity.value</td>
<td>93</td>
<td>0.974</td>
<td>0.835</td>
<td>0.899</td>
</tr>
<tr>
<td>Dosage.timing.repeat.when</td>
<td>21</td>
<td>1</td>
<td>0.571</td>
<td>0.727</td>
</tr>
<tr>
<td>Dosage.asNeededBoolean</td>
<td>38</td>
<td>0.913</td>
<td>0.583</td>
<td>0.712</td>
</tr>
</tbody>
</table>

Discussion

In this study, we proposed an integration framework to support the representation of structured and unstructured EHR data into the FHIR model, leveraging both the NLP-based mapping rules and structured data ETL methods. The FHIR-based type system was used as the integration central-bus for the interchange between multiple information extraction tools. We performed the initial experiment and evaluation on the FHIR resource “MedicationStatement”.

As the FHIR model is very rich and comprehensive with a multi-layered structures and elaborate definitions, in spite of creating mappings rules between NLP output types and FHIR model elements through delicate model analysis, more model elements are still being studied and the current mapping rules are far from complete. However, our results are effective in proving the feasibility of leveraging NLP tools to support the conversion of EHR data into the FHIR model. Currently, in our use case, MedXN and MedTime are used to perform the extraction of medication and temporal entities and attributes from medication lists. In the future, particular entity extraction modules will be developed for the reminder NLP extracted entries from medications list.

Our experiment was implemented using the UIMA architecture, in which we created an FHIR-based type system. We demonstrated that the FHIR type system was effective in facilitating integration of the differing outputs of multiple UIMA-based NLP tools. This mechanism provides powerful extension capability in that other UIMA-based NLP tools not part of this study can be easily integrated. In fact, we are working on the extension implementation on other FHIR resources representation. As part of the next step, we plan to adopt a clinical NLP tool cTAKES to acquire relevant NLP output for the FHIR resource types, such as Condition, Procedure, and Diagnostic Report. The cTAKES type system, based on Intermountain Healthcare’s Clinical Element Models (CEM), is currently being studied and mapped into the FHIR elements.

One of ongoing challenges is that the standardization work in FHIR is still in progress: the “code system” list and the “value set” list construction are on progressing, and more standardization work needs to be done as part of our implementation. These include normalizing Dosage.route using the defined FHIR value set [http://hl7.org/fhir/ValueSet/route-codes], and integrating the standardization service into our pipeline.

Currently, we manually downloaded the FHIR resource in turtle format, and then extracted the resources classes and properties to build an FHIR schema for annotation. In a future design, we plan to automatically generate the FHIR annotation schema using the FHIR RDF specification.

Although the performance of NLP has room for improvement due to the complexity of clinical notes, our mapping rules relying on NLP output were able to populate FHIR elements instances into most elements defined in the FHIR MedicationStatement model. Our FHIR type system-driven integration method provides a generic and scalable framework to support the FHIR modeling of structured and unstructured data.
Conclusion

In this study, we developed a clinical NLP pipeline using an FHIR type system for integrating structured and unstructured EHR data. We demonstrated the feasibility of our approach, focusing on the core elements of the FHIR resource “MedicationStatement”. We are actively working on creating mapping rules and annotation work on other FHIR resource types and on improving the performance of our integration approach. We believe that the standard FHIR modeling method for EHR data, as illustrated in this study, will benefit future semantic data exchange and integration, and rapid generation of computational phenotypes for advancing clinical and translational research.

Acknowledgements

This study is supported in part by NIH grant U01 HG009450.

References

Usability Evaluation of an Unstructured Clinical Document Query Tool for Researchers

Gretchen Hultman, MPH¹, Reed McEwan, MS, MSSE², Serguei Pakhomov, PhD³, Elizabeth Lindemann⁴, Steven Skube, MD⁴, Genevieve B. Melton, MD, PhD¹,⁴
¹Institute for Health Informatics, ²Academic Health Center – Information Systems, ³College of Pharmacy, ⁴Department of Surgery, University of Minnesota, Minneapolis, Minnesota

Abstract

Natural Language Processing – Patient Information Extraction for Researchers (NLP-PIER) was developed for clinical researchers for self-service Natural Language Processing (NLP) queries with clinical notes. This study was to conduct a user-centered analysis with clinical researchers to gain insight into NLP-PIER’s usability and to gain an understanding of the needs of clinical researchers when using an application for searching clinical notes. Clinical researcher participants (n=11) completed tasks using the system’s two existing search interfaces and completed a set of surveys and an exit interview. Quantitative data including time on task, task completion rate, and survey responses were collected. Interviews were analyzed qualitatively. Survey scores, time on task and task completion proportions varied widely. Qualitative analysis indicated that participants found the system to be useful and usable in specific projects. This study identified several usability challenges and our findings will guide the improvement of NLP-PIER’s interfaces.

Introduction

Optimal requirements and design considerations for a Natural Language Processing (NLP) tool intended for use by clinical researchers are not well understood. Achieving a high degree of system usability is key to creating a product that is well accepted by target users. Usability is defined as “the effectiveness, efficiency, and satisfaction with which the intended users can achieve their tasks in the intended context of product use”¹. Usability testing is employed to help ensure that an end product is usable. Usability testing is considered an effective usability methodology that has high strategic impact². It can be defined as “the process of learning about users by observing them using a product to accomplish specific goals important to them”³. Testing typically involves representative users performing realistic tasks under typical conditions³. It is a component of user-centered designed processes that is employed to create products with high usability.

Several studies have examined usability challenges associated with search engines. One study by Dudek, Mastoram and Landoni examined the importance of usability in the evaluation of general search engines⁴. Participants were asked to perform tasks using a search engine. This study found that users valued usability when selecting a search engine⁴. Kushniruk et al. conducted an evaluation of a new automated text summarization system called Centrifuse that was designed to enable patients and families to search for information about health conditions⁵. They also evaluated three existing search engines⁵. Think-aloud data from this study revealed that users had no clear preference for one system over another but liked different aspects of different systems⁵. Trivedi et al. evaluated NLPReViz, a tool designed to allow clinical researchers to train, revise and revise NLP models for use on clinical text⁶. Clinicians were given a tour of the interface and an introduction to the tool, then were asked to build models and discuss their experience with this tool⁶. This study found that physicians were able to use the tool to build models and provided generally positive feedback⁶.

Natural Language Processing - Patient Information Extraction for Researchers (NLP-PIER) has been previously described and is a clinical notes processing platform including an NLP query and search engine for clinical and translational researchers developed at the University of Minnesota⁷. For the purpose of this study, the user-facing application components of NLP-PIER comprise the “system” being evaluated. This system (search engine + graphical user interface) is a three-tiered web application enabling users to submit queries to an application server that transforms user input into search requests against an Elasticsearch backend holding approximately 100 million indexed clinical notes from the enterprise Fairview Health system EHR. The system is secured by an authentication and authorization layer in the application server that ensures authenticated users only have access to sets of notes that are defined externally and configured in the Elasticsearch engine. This system was designed to give clinical
researchers access to NLP capabilities for searching clinical notes in an environment that is compliant for accessing protected health information (PHI). Unstructured data contains valuable information but is difficult to “unlock” for automated secondary uses. This system fills a need by making unstructured clinical text more usable for researchers. We developed the NLP-PIER system to enable “everyday” clinical researchers to search clinical notes. Similar systems have been deployed at other institutions.

NLP-PIER has two search modes: full text searching (Figure 1) and concept searching (Figure 2 and Figure 3). We anticipate the following workflow for this application: First, the user identifies a research project in which clinical notes are needed for research. Then, the user would request access to the relevant set of notes from our institution’s Clinical and Translational Science Institute (CTSI). Participants would gain access to the note set and the application after approval. Users would be granted access to the search engine that fit their needs and would switch between the two interfaces after authenticating against an identity management system. Next, NLP-PIER is used to search through the data set, likely starting with broad searches and narrowing to more useful search terms. Lastly, users export the results out of the system for further research. Often, this research may involve using the data to compile a list of patients for chart review. Figure 2 shows what search results look like when they are returned by the system. We anticipate that this system would be most useful for research projects with data that is not easily found in structured fields. For example, searching for patients with terminal ileitis or searching for patients presenting to the emergency department with abdominal pain (tasks that require access to information hidden in the unstructured part of the EHR), but not appendicitis (easily found through diagnosis codes).

**Full Text Search**

This search mode relies on standard information retrieval techniques for indexed content and functions similar to search engines with which computer users are familiar. The raw text of clinical notes from the EHR were indexed for keywords by Elasticsearch using the Snowball Analyzer. The full text search interface accepts text queries using the Lucene query syntax for keywords and phrases. Parenthetical grouping of search terms and logical operators (AND, OR, NOT) allow the construction of more complex queries than the default operand (AND) assumed between terms. Sorted and paged results are returned based on a standard term frequency/inverse document frequency (TF/IDF) score for the search term(s), with the highest scoring notes listed first. We choose to pursue this design because it is familiar to users who use other search applications. Search results are presented in a list with a section of text present and a link to metadata for the note in the corner (Figure 2).
Search terms in the full text interface are often instances of a medical concept: either a common synonym or a single, specific term in a controlled vocabulary. Because concepts can be represented by multiple terms (strings) in the search corpus, any single search term, or even a small set of synonyms, used in a full text search will likely fail to match documents where the concept is expressed using terms not in the search input. The United Medical Language System (UMLS) meta-thesaurus solves the problem of multiple expressions for a single concept by mapping terms from disparate vocabularies to unique concepts represented by a Concept Unique Identifier, or CUI\(^\text{13}\). Because multiple terms from multiple vocabularies roll up to a single concept, searching by CUI instead of specific lexical variant amounts to a form of query expansion in the full text search sense. This form of query expansion is the basis of the conceptual search interface in NLP-PIER. CUI search functionality is enabled in NLP-PIER by indexing CUIs, identified with the UIMA-based\(^\text{14}\) BioMedICUS\(^\text{15}\) NLP system through which each clinical note is processed, with the note text.

The search box in the conceptual search interface (Figure 3) gives the user the ability to search for clinical notes containing identified CUIs using a free text, auto complete input box. As the user types in text, NLP-PIER looks for free text matches drawn from the UMLS and suggests matching terms. Suggestions can be constrained by selecting a specific UMLS vocabulary and/or semantic type from the Specialist’s Lexicon\(^\text{16}\). When the user selects a suggested term, NLP-PIER displays and keeps track of the CUI so it can be used in a query against the indexed CUIs. The interface supports combining CUIs using logical operators and specifying whether the concept is used in a negated context (Figure 4). Results are displayed in the same way as for the full text interface. When making design choices for this interface, it was important to account for the experience of end users. We anticipated, based on anecdotal evidence, that users would not be familiar with this type of searching. We attempted to simplify the design for the end user and borrowed elements from designs (such as web applications) that are likely to be familiar to these users.
Objectives

The purpose of this study was to evaluate the usability and user acceptance of the NLP-PIER system through direct user testing to gain insight into interface design opportunities, user acceptance, and user preferences with the platform’s two main modes: full text searching and UMLS concept searching. These findings represent a step forward toward understanding the functionality and usability needs of clinical researchers when using a self-service NLP tool for searching clinical notes.

Methods

This study was conducted at the University of Minnesota within the Academic Health Center – Information Exchange and its associated secure data shelter which serve as key components of NLP-PIER’s infrastructure enabling clinical research in a Protected Health Information (PHI)-compliant environment.

Tasks

In order to control for variability in the use of NLP-PIER we constructed an experimental protocol designed to be a set of realistic tasks relevant to clinical researchers. Tasks were designed to identify cohorts of patients that could not easily be identified using structured data. This choice was made because NLP-PIER was designed to provide distinct and added value for information only found in unstructured format. Two physicians reviewed the protocol of constructed tasks for clinical accuracy. A software developer reviewed tasks for technical accuracy. High level examples of the tasks included: performing a search in the full text interface, running a query in the concept search interface, filtering search results to find notes written by providers, and exporting the results of queries to an external program. Table 1 contains all tasks performed in each interface.

Table 1. List of Tasks.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td>Search &quot;chest pain&quot;</td>
</tr>
<tr>
<td>Search</td>
<td>Restrict search to notes written by providers</td>
</tr>
<tr>
<td>Search</td>
<td>Remove administrative notes</td>
</tr>
<tr>
<td>Search</td>
<td>Save the search</td>
</tr>
<tr>
<td>Search</td>
<td>Restrict to 2015</td>
</tr>
<tr>
<td>Search</td>
<td>Rerun saved query</td>
</tr>
</tbody>
</table>
A convenience sample of eleven clinical research faculty participated in this study. Participants were recruited by identifying clinicians who engaged in research and sending an invitation to participate via email. Eleven participants responded and agreed to participate. Data collection sessions took place in the participant’s office and lasted approximately 30-45 minutes. For all sessions, participants were seated at a laptop computer. Screen capture software (Voila!) was used to record sessions. Participants were logged into NLP-PIER and were given a tour of the interface. Each participant was given an opportunity to ask questions. Participants were also provided with a “tip sheet” of helpful hints. This was done to partially mimic “real-world” setting, where participants would have time to test out the interface and ask questions of colleagues. The first part of the usability assessment consisted of each participant completing two sets of tasks in the full text search interface. Following completion of tasks, participants completed the SUS and raw NASA-TLX surveys.

The SUS is a widely used tool that was designed to be an efficient assessment of self reported usability. It was designed in response to a need to establish a broad measure of usability that could be used to compare usability across different contexts. The SUS is “a ten item scale giving a global view of subjective assessments of usability”. The SUS score ranges from 0 to 100 with a higher score indicating a more usable system. We selected this tool because it is a standardized and widely accepted tool for measuring self-reported subjective assessment of usability.

The raw NASA-TLX is a widely used scale designed to measure subjective impressions for workload. The NASA-TLX consists of six subscales across different domains of workload: 1. Mental; 2. Physical; 3. Temporal Demands; 4. Frustration; 5. Effort; and 6. Performance. The assumption of the NASA-TLX is that the combination of these six scales represents the abstract idea of “workload”. Participants rate workload on a scale with 21 gradations for each domain. A higher score represents a higher workload. We used this tool because it is a standardized way of measuring the subjective workload a participant experiences when using an application.

Part two of the assessment was completed in the concept search interface. Similarly, participants completed two different sets of tasks using this interface. Participants then completed the SUS and NASA-TLX surveys as well as a demographic survey and a brief interview. To analyze participant opinions and feedback, we conducted interviews with participants. In post-test interviews, participants were asked the following questions:
1a) How useful do you find the NLP-PIER system?
1b) Do you have any suggestions for making it more useful?
2a) How easy do you think NLP-PIER was to use?
2b) Do you have any suggestions for making it easier?
3) Do you have any current or future projects in which you could envision using NLP-PIER?
4) Is there anything else we should know about your experience using NLP-PIER?

Analysis

To measure the different domains of usability, we analyzed our data along five main axes. In order to measure satisfaction, we analyzed SUS scores. Efficiency was measured by the time spent on task. We measured effectiveness by analyzing task completion or reason for lack of completion. We also analyzed workload through the NASA-TLX. Lastly, we compiled participant comments and feedback. To analyze time on task and task completion, we conducted content analysis on the video recordings of the sessions. For each task, we coded the start time, end time, whether the participant successfully completed the task or not, and the reason lack of completion if applicable. Tasks were coded as either competed or failed and the reason for the decision was recorded. For each task, we performed cognitive task analysis to identify the sub-tasks involved in the task. This helped us to understand the reason for a participant failing to complete a task. Two coders (RM and GH) coded two videos (18%) and discussed all discrepancies in order to establish a standard process. Questions were discussed and resolved between the two coders. One coder was a developer and the other was an informatics researcher with experience in public health. One coder (GH) coded the remaining videos. Any outstanding questions or issues were resolved between the two coders. Interviews (n=10) were recorded, transcribed and coded using QSR International’s NVIVO 11 software. One interview was excluded due to lack of completion. Two coders (EL and GH) coded all interviews (Cohen’s kappa 0.84, percent agreement 98%). Both coders were informatics researchers with experience in qualitative analysis. For each interview question, we coded the response and any associated comments made by the participant.

Results

Participants, all of whom were practicing physicians, represented a variety of different specialties including colorectal surgery, gastroenterology, hematology, and plastic surgery. Of participants who responded, 6 were male and 4 were female. Four participants reported having practiced for less than five years post-residency, three reported practicing for five to ten years, and two reported more than ten years of practice experience post-residency. All participants were medical doctors. Participants spent an average of 15 hours per week (range 7 to 30 hours) on research activities and all reported being “average” users of technology. Two-thirds had experience requesting data from our institution’s clinical data repository. Participants were not associated with the Health Informatics Department and did not have prior knowledge about NLP-PIER.

Survey Data and Content Analysis

For the full text search interface, the SUS score was 69.4 (SD 19.8), and for the concept search interface the SUS score was 66.1 (SD 32.4). The NASA-TLX score for the full text search interface was 18.8 (SD 5.7), and the NASA-TLX score for the concept search interface was 21.8 (SD 7.7) (Table 2). Content analysis revealed wide variation in time on task (1.4 sec – 85.3 sec), and task completion percentage (9%, n=1 to 100%, n=11). Because this study represented a first step toward understanding the usability of NLP-PIER, we did not have initial estimates of how long certain tasks would take and estimated that some tasks would be faster than others. To add context, we determined what tasks participants did not complete successfully and why they failed to complete the task. We determined this by performing cognitive task analysis and identifying sub-tasks of each task. Tasks were of varying difficulty, consisting of between 1 and 7 subtasks. We noted where in the task the participant deviated from the expected path. Table 3 summarizes tasks that were completed by less than 50% of participants and reasons for lack of completion. Table 4 summarizes tasks that had the highest completion percentage.
Table 2. Least completed tasks.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Task</th>
<th>Mean Time in Seconds (stdev)</th>
<th>Task completion proportion</th>
<th>Subtasks</th>
<th>Reasons for Lack of Completion</th>
</tr>
</thead>
</table>
| Full text | Restrict search to notes written in 2015 | 85.3 (57.8)                 | 1/11                       | 1. Go to help menu  
2. Find date query  
3. Click add example to search box  
4. Make modifications  
5. Click search | Participants (n=10) did not go to help menu to use a pre-made query in the help menu to restrict by date. |
| Full text | Re-run saved query                        | 31.4 (29.9)                 | 2/11                       | 1. Click arrow next to all notes  
2. Click on saved query | 1 participant did not save the query  
8 did not click arrow to open menu where saved queries were |
| Full text | Clear a previous search                   | 6.3 (7.6)                   | 3/11                       | 1. Clear text  
2. Clear filters (alternatively, participants could refresh the webpage) | Participants (n=8) did not clear filters |
| Concept | Run a query                               | 21.0 (32.8)                 | 5/11                       | 1. Click add to query  
2. Click search | 4 participants did not complete previous step  
2 did not click search |

Table 3. Most completed tasks.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Task</th>
<th>Mean time in seconds</th>
<th>Task completion proportion</th>
<th>Subtasks</th>
<th>Reasons for Lack of Completion</th>
</tr>
</thead>
</table>
| Full text | Switch number of results displayed per page | 17.0 (12.4)          | 9/11                       | 1. Go to box next to search  
2. Click down arrow  
3. Pick 20 from list  
4. Click search | 2 could not change results because they did not refresh results in previous step |
| Full text | Export search results                     | 12.7 (20.8)          | 9/11                       | 1. Click export button  
2. Click button on pop-up | 2 could not export results because they failed to refresh results in previous step |
| Full text | Open exported file                        | 17.5 (19.1)          | 9/11                       | 1. Doc will automatically open in excel  
(can perform additional steps to change options) | 2 could not open exported file because they failed to export results |
| Concept | Search any vocabulary                     | 4.7 (6.4)            | 10/11                      | 1. Use default options | 1 used a different option from the menu |
| Concept | Select sign or symptom                    | 18.7 (10.0)          | 11/11                      | 1. Use default options | N/A |
| Concept | Use any vocabulary                        | 1.4 (3.6)            | 10/10                      | 1. Use default options | One skipped task |

Qualitative Results

In interviews, all participants expressed that NLP-PIER was easy to use. One stated “very easy, I don’t have any suggestions I think it’s very straight forward.” Two stated that the full text search engine was easier to use than the
concept search interface with one stating “I thought number one (full text searching) was super simple to do the search and number 2 (concept searching) was obviously more challenging but as I said before I would welcome that challenge if it was more precise”. All participants stated that NLP-PIER would be useful with one stating “I think it could be really useful”. Two participants expressed reservations about using the system in its current state but noted that it had potential to be useful. All participants had examples of specific clinical research and quality assurance projects in which NLP-PIER would be useful. Examples included: locating patients who present to the emergency department with facial weakness and locating pediatric patients with intestinal inflammation/colitis. Despite positive feedback, users had several suggestions for improving NLP-PIER. For concept searching, participants brought up concerns about the two-step process for running a query and understanding how negation options functioned. One participant expressed concerns about not understanding how the search engine was functioning. For the full text search mode, participants had concerns about how to run advanced queries, how to search for negated terms, and how the interface would handle misspellings. For both modes, participants had concerns about the amount of text that appears from each note in the search results, were confused about the process of refreshing results, and were confused about creating filters. Additionally, participants had questions about exporting the search results and utilizing them in future research activities. They also expressed a desire to be able to search on other types of reports such as laboratory and imaging reports.

Discussion

Our usability experiment with formal user testing of NLP-PIER provides a valuable evaluation of a self-service tool for clinical researchers with NLP capabilities for EHR notes. Survey scores were similar across the two interfaces. Because SUS is judged on a scale from 0-100, it can be difficult to interpret what the score means. Bangor et al. suggested interpreting scores on a standard letter grade scale where scores below 70 indicated that the product had substantial usability issues. SUS scores for both interfaces indicate that interfaces have marginal usability or a “D” on a letter grade scale. This demonstrates the need to improve usability before further deploying the system. Interestingly, participants verbally indicated that they thought NLP-PIER was usable. Further work should explore this discrepancy. While we initially spoke with researchers about their interest in the system, more could have been done prior to development to understand the needs of researchers. This includes conducting a needs assessment, workflow analysis, or interviews with participants to better understand the needs of researchers before and during the process of creating this system. Large inter-participant variability was observed for both SUS and NASA-TLX scores. This underscores the difficulty in creating a tool that is useful to a wide variety of clinicians. Factors such as age, gender, and computer experience can all affect usability and tools like NLP-PIER need to be useful to researchers with a wide variety of backgrounds and experiences.

Our usability evaluation illustrated several opportunities to improve the system for clinical researchers. We brainstormed solutions to the challenges that we identified. Table 4 summarizes challenges and proposed solutions. Despite these challenges, all participants in this study thought NLP-PIER would be useful in their research and all indicated that there were current or future projects in which NLP-PIER would be useful. This aligns with other work that suggests that this type of system is useful in helping researchers harness the data in unstructured clinical notes for secondary purposes such as research and quality improvement projects. Importantly, we did not evaluate the results of user queries for completeness. Further evaluation is necessary in the future. Also, future work should focus on evaluating the existing functionality and identifying additional functionality that would be useful for researchers.
Table 4. Summary of overall usability challenges and proposed solutions.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Proposed Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of ability to notice when results need to be refreshed</td>
<td>Clear the previous results or overlay them with a semi-transparent background on which the need to refresh is indicated</td>
</tr>
<tr>
<td>Wanting number of patients to be displayed at the top of search instead of number of notes</td>
<td>Add optional unique patient count calculation toggle</td>
</tr>
<tr>
<td>Wanting to being able to export list of unique patients</td>
<td>Will not implement at this time, this can be done with post export data manipulation</td>
</tr>
<tr>
<td>Overly complex filters and unclear filter titles</td>
<td>Group filters into EHR and HL7-LOINC categories and add mouse over elements exposing definitions of each data element used as a filter</td>
</tr>
<tr>
<td>Not clearing filters when attempting to clear a previous query</td>
<td>When user submits a query without clearing or modifying previous query’s filters, alert user that filters will be re-used</td>
</tr>
<tr>
<td>Wanting number of notes displayed more prominently</td>
<td>Increase font size and move to be more prominent</td>
</tr>
<tr>
<td>Formatting of displayed note text</td>
<td>Use formatted plain text parsed from rich text input</td>
</tr>
</tbody>
</table>

Our study revealed several findings related to the general needs of clinical researchers when using NLP tools in research. First, our research indicated that participants are interested in these types of tools and find them to be applicable to their work. Secondly, our research revealed several usability challenges specific to the concept search interface. Our results indicated that this type of searching presented a number of challenges and seemed unfamiliar to our participants. Future work should be done to explore clinicians’ understanding of this type of system. It is likely that many researchers require additional training before feeling comfortable using this system, and may prefer full text searching when it is sufficient to meet their needs. While not part of the formal interview, several participants mentioned informally that the duty of using the tool would likely not fall to the principal investigator but to research assistants and students. Therefore, it is necessary to make tools usable for this population of users. Our study had a number of limitations. We had small a sample size and recruited participants from one user group, thus limiting the generalizability of our findings. Additionally, this study was conducted at a single health system and may not be generalizable to other institutions.

Conclusion

This study sought to employ usability testing to evaluate a self-service search engine for its usability with clinical researchers. We identified a number of usability challenges and future mechanisms to improve NLP-PIER to address the concerns of users. Additionally, we identified barriers related to the experience and familiarity around concept searching and the system’s associated interface for end users. This study also demonstrated that substantial variation exists between different users. At the broadest level, our findings illustrate the importance of incorporating user testing and feedback in the system design process.

Acknowledgements

We would like to acknowledge the Fairview Health Services Data Reporting and Analytics team UMN Academic Health Center-Information Services, the Natural Language Processing-Information Extraction Program, the Agency for Healthcare Research & Quality (#R01HS022085 (GM)), and National Institutes of Health (#R01LM011364 (GM), #R01GM102282 (SP) and #8UL1TR000114 (Blazer)).
References

20. NVivo qualitative data analysis Software. QSR International Pty Ltd.; 2015.
CDISC SHARE, a Global, Cloud-based Resource of Machine-Readable CDISC Standards for Clinical and Translational Research

Samuel Hume, DSc 1, Anthony Chow1, Julie Evans2, Frederik Malfait3, Julie Chason1, J. Darcy Wold1, Wayne Kubick, MBA4, Lauren B. Becnel, PhD 1,5
1Clinical Data Interchange Standards Consortium (CDISC), Austin, TX, USA; 2Samvit Solutions, LLC, Reston, VA, USA; 3TransCelerate Biopharma, Conshohocken, PA, USA; 4Health Level7 (HL7) International, Ann Arbor, MI, USA; 5Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA

Abstract

The Clinical Data Interchange Standards Consortium (CDISC) is a global non-profit standards development organization that creates consensus-based standards for clinical and translational research. Several of these standards are now required by regulators for electronic submissions of regulated clinical trials’ data and by government funding agencies. These standards are free and open, available for download on the CDISC Website as PDFs. While these documents are human readable, they are not amenable to ready use by electronic systems. CDISC launched the CDISC Shared Health And Research Electronic library (SHARE) to provide the standards metadata in machine-readable formats to facilitate the automated management and implementation of the standards. This paper describes how CDISC SHARE’s standards can facilitate collecting, aggregating and analyzing standardized data from early design to end analysis; and its role as a central resource providing information systems with metadata that drives process automation including study setup and data pipelining.

Introduction

Billions of dollars are spent on clinical and translational research annually. Due in part to the deficient application of data standards, data from these studies exist in non-interoperable silos. Siloed, non-standard data cannot be easily reused for meta-analyses, hypothesis generation, safety signal detection, or other secondary uses. The Clinical Data Interchange Standards Consortium (CDISC) is a non-profit standards development organization (SDO) that develops and supports global, vendor-neutral, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. For 20 years, CDISC and its community of thousands of volunteers from over 435 member organizations and the research community at large has developed standards for collecting, tabulating, exchanging, submitting and archiving clinical and translational research, including clinical trials as well as preclinical, nutrition, comparative effectiveness, epidemiology, safety surveillance and other public health research. Each standard and its implementation guide contains normative content, which are the prescriptive elements that describe the scope and the metadata that comprise the standards, and informative content that helps implementers to understand how to apply the standard in their organization. These standards are developed and maintained using robust processes that foster stakeholder engagement, community comment, governance and maximal transparency. CDISC standards have been downloaded in approximately 100 countries and are indexed within both BioPortal1 and FAIRsharing.org (formerly BioSharing.org)2.

Beginning December 2016, both the US Food and Drug Administration (FDA)3, 4 and Japan’s Pharmaceuticals and Medical Device Agency (PMDA)5, 6 began requiring new trials’ submissions to comply with certain CDISC standards. These standards are recommended by the European Medicines Agency (EMA) Clinical Trial Advisory Group7 and the Chinese FDA8. CDISC also has alliances across the National Institutes of Health (NIH) including a long-standing collaboration with National Cancer Institute (NCI) Enterprise Vocabulary Services (EVS) team9, National Institute for Allergies and Infectious Diseases (NIAID), which has used CDISC standards for pharmacovigilance studies and meta-analyses and requires them for new HIV research protocols and in Brazil, Russia, India, China and South Africa (BRICS) studies; the National Institute for Diabetes and Digestive and Kidney Disorders (NIDDK), which was part of the consortium that developed CDISC’s polycystic kidney disease TA; and others.

CDISC’s standards are open and freely available via the CDISC Website10 as PDFs. These files are human readable, but not easily computable by software, so implementers had to extract normative metadata from each standard to incorporate it into their CDISC-compliant information systems. To obviate this issue, CDISC created the CDISC Shared Health and Research Electronic Library (SHARE), a metadata management platform that includes a metadata repository (MDR) and ecosystem of tools and services that, together, support standards development teams in creating
content and standards implementers (e.g., researchers, industry) by making its standards available in a variety of electronic formats including RDF, XML, JSON, CSV, and PDF.

Here we describe CDISC SHARE, including methods to retrieve machine-readable standards metadata and examples of real world applications.

Materials and Methods

Standards Development Process. CDISC standards are internally and publicly reviewed through a standardized operational process11 where releases are made available on the CDISC public Atlassian Confluence Wiki website for comment during a specified interval of time. Any member of the public may create an account in the CDISC Wiki and mark up candidate release documents with comments that are automatically integrated into CDISC’s instance of JIRA. Comments are then compiled and reviewed, and changes are made to the model where persuasive. The final model, its documentation and all comments (along with their disposition/resolution) are made available to the public. For controlled terminology, approved packages are managed within NCI’s EVS teams’ systems, where their curators perform quality control and annotate the terminologies with links to synonyms and related terms indexed within EVS. The CDISC SHARE team extracts these annotated, curated terminologies through an automated pipeline where the content is imported into the CDISC SHARE MDR, and the CDISC SHARE biocurators assess post-import logs for any failures or other issues that require resolution prior to final publication in the MDR.

Metadata Repository (MDR). CDISC SHARE provides a cloud-based CDISC standards specific MDR for the curation, management, and publication of the standards metadata in machine-readable formats. CDISC SHARE is currently comprised of a commercial Java Enterprise Edition application backed by an Oracle database, running in JBoss middleware. The MDR implements the Object Management Group’s Reusable Asset Specification12 as a base model upon which the SHARE metamodel is layered. The CDISC SHARE metamodel is based on the International Standards Organization’s (ISO) standard for metadata registries, ISO/IEC 1117913, and forms the foundational basis to structure information about metadata assets, their relationships, and versioning. The CDISC SHARE RESTful Application Programming Interface (API) provides an interface for software applications to query and retrieve standards metadata from the MDR. For implementers not using CDISC SHARE API-enabled software, CDISC SHARE provides an export feature based on the API libraries that enables the CDISC SHARE biocurators to publish full standards for download by implementers. Implementers download the content from the CDISC Website10 in several formats including XML, RDF, and CSV; and manually import the files into local metadata management systems for use by downstream research information systems.

CDISC SHARE Software Ecosystem. The CDISC SHARE Ecosystem is a software tools platform that complements the CDISC SHARE MDR by extending additional services to standards developers and implementers. The CDISC SHARE Ecosystem includes CDISC’s instance of Atlassian Confluence Wiki for active standards development and publication support; JIRA for issue tracking and community public review; Bitbucket for collaborative development and version control; and numerous bespoke software applications for loading and extracting MDR content.

Quality Control (QC) Pipeline. After internal and public reviews are completed, final standards versions enter the CDISC SHARE QC pipeline. Normative content is extracted from the standards and assessed by expert human curators. A variety of QC scripts are then run against the metadata to ensure that datatypes, character limitations, and other quality items are appropriate. Where issues are uncovered they are corrected with support from standards development teams, as required. QC’ed metadata are then imported into the CDISC SHARE metadata repository, where a second set of scripts are run against the data. These scripts enforce CDISC SHARE model rules of referential integrity identifying cross-standard issues as well as violations of CDISC SHARE versioning rules. Once loaded, content is exported and compared to the original import content to ensure correctness.

Communications. As new standards are imported into the MDR and curated, CDISC Communications notifies the public by updating the CDISC SHARE Exports webpage, releasing notifications in the CDISC Newsletter, sending email blasts, posting updates to CDISC’s LinkedIn groups, and tweeting from the CDISC SHARE Twitter account. CDISC SHARE Workshops and CDISC SHARE Showcases that highlight innovative software solutions that implement CDISC SHARE content are communicated via these same channels. The public can register for the Newsletter, public webinars, and International CDISC Conferences on the CDISC Website10.

Results

CDISC Standards Organization. CDISC standards (Figure 1) are categorized as Foundational Standards, Semantics, Therapeutic Area (TA) Standards, and Transport Standards. Foundational Standards include Standard for Exchange
of Nonclinical Data (SEND)\textsuperscript{14} for the collection and tabulation of animal model and other pre-clinical data, Protocol Representation Model (PRM)\textsuperscript{15}, Analysis Data Model (ADaM)\textsuperscript{16} for defining analysis datasets, Clinical Data Acquisition Standards Harmonization (CDASH)\textsuperscript{17} that provide a minimal set of data elements common to essentially all studies, the Study Data Tabulation Model (SDTM)\textsuperscript{18} for data tabulation, and others.

![End to End Standards & Process Automation](image)

**Figure 1.** CDISC End-to-end Standards including Foundational Standards, Semantics, and TA Standards

Except for Protocol, the foundational standards and their implementation guides are organized into classes of related categories of information. Within SDTM for example, each class is subdivided into domains, and each domain is comprised of variables. Semantics include CDISC Controlled Terminologies, developed in collaboration with NCI EVS, and the Biomedical Research Integrated Domains Group (BRIDG) domain information model. TA User Guides provide indication-specific “slices” of foundational standards and their terminologies, with examples of how to use these standards for a given TA. Transport Standards support submissions, data interchange, study archival, and automated study setup\textsuperscript{19}.

CDISC implementation and user guides contain normative and informative content. Normative content includes the classes, domains, variables, definitions, and other core metadata of a standard. Informative content includes examples, assumptions and other information that help adopters to determine how to best implement the standard within their organization. While tools within the CDISC SHARE Ecosystem such as the Confluence Wiki and JIRA are utilized to create and maintain both normative and informative content, the CDISC SHARE MDR contains the normative standard metadata to avoid confusion of an actual standard versus examples of implementation.

While each CDISC standard represents a different aspect of the clinical research data lifecycle and is developed as an independent standard, the CDISC standards are also intended to work together across the lifecycle. When the standards are organized to represent the flow of information across the clinical research data lifecycle this is called the end-to-end standards representation.

**CDISC SHARE Model Version 1.0.** Normative CDISC standards metadata in the MDR conform to a version 1.0 CDISC SHARE model. SHARE content is organized into layers of increasingly concrete models, starting with the Open Management Group (OMG) Reusable Asset Specification (RAS) as the foundational model, next implementing International Standards Organization (ISO)/International Electrotechnical Commission (IEC) 11179 as the metamodel, and then adding a CDISC standards model as top-layer. ISO/IEC 11179 forms the foundational basis of CDISC SHARE’s metamodel to structure information about metadata assets, versioning, ownership and other core metadata information. RAS and ISO/IEC 11179 are standards commonly implemented in metadata management systems, but the CDISC model is novel as it is the first time CDISC has structured their contents standards into a one common model.

The CDISC SHARE metamodel maintains the relationships needed to represent the end-to-end CDISC model that captures how one standard flows into the next in the clinical research data lifecycle. These are relationships that exist between each of the individual foundational standards. For example, a CDASH variable may include a maps-to relationship to an SDTM variable. This is a two-way relationship that also provides a maps-from relationship linking the SDTM variable back to the CDASH variable. The CDISC SHARE end-to-end standards model represents the first time the standards have been explicitly represented as one full-lifecycle standard as opposed to distinct standards intended to work together. To represent the model adequately, relationships implicit in the CDISC foundational standards have been made explicit in the model. The number of relationships currently maintained in CDISC SHARE exceeds the number of metadata elements by a ratio of approximately 5:1. The end-to-end standards representation provides support for standards metadata-driven automated data pipelining.

The CDISC model is organized within SHARE’s metamodel as Metadata Elements (MDEs), which are discrete pieces of metadata such as an individual variable, and Metadata Element Sets (MDESs), which are categories of MDEs such
as a domain or groupings of MDESs such as a class that categorizes multiple domains. The metamodel also represents semantics which are organized into value domains that are in turn comprised of one or more domain values. These metamodel elements are used to represent the code lists and individual terms found in the CDISC Controlled Terminology.

Within each foundational standard the metadata are represented in the SHARE metamodel as a hierarchy of related components rooted in the parent top-level MDES (Figure 2). This MDES contains both class and domain MDESs. Class MDESs can include both domain MDESs and their own variable MDEs, though most variable MDEs are children of the domain MDESs. Variables often have an associated CDISC Controlled Terminology code list captured by the *represented-by* relationship.

**Figure 2.** CDISC SHARE metamodel hierarchy

**CDISC SHARE Application Programming Interface (API).** The CDISC SHARE metadata are exposed for retrieval through a RESTful API. The API’s schema and other documentation are available on the CDISC Website\(^\text{20}\), and this schema’s endpoints represent CDISC SHARE metamodel MDEs and MDESs (Figure 2). Each CDISC SHARE metadata object has a unique object identifier (OID). API end users can request an entire standard, including all of its domains, variables and URLs to retrieve the latest terms for each code list. Alternatively, the API service recursively traverses this standards hierarchy to fulfill requests for content at various levels of granularity. To request the metadata for a specific domain, for example, an end user “drills down” through the standards hierarchy to that domain, first identifying the OID for a specific standard, then the OID for the domain’s parent class, and finally the OID for the domain itself. The API supports discovery by providing the resources to retrieve either the next level of detail within the standards hierarchy or the full top-level standard.

Several optional query parameters exist for the API’s endpoints. For full, top-level standards such as the SDTM Model and Implementation Guide, end users can specify the returned media type as either XML or RDF-XML, Turtle or N-Triples. For components of these standards, either XML or JSON may be requested. API users can specify whether they would like to retrieve standards with a lifecycle status of Approved Final or Approved Provisional. Finally, end users can request that returned metadata remain uncompressed or be compressed using gzip or zip. Details on attaining access to the CDISC SHARE API are available on the CDISC Website\(^\text{10}\).

The availability of a standards MDR containing metadata representing the full clinical research data lifecycle with an API available to support process automation is uncommon among biomedical research and healthcare data standards. **Table 1** lists well-known data standards applied to healthcare and methods of accessing the standards metadata, including whether the standards are consumable via an API.
# Table 1. Comparison of Methods for Accessing Standards Metadata (see Table 2 for acronym definitions)

<table>
<thead>
<tr>
<th>Standard</th>
<th>Methods of Accessing Computable Standards Metadata</th>
<th>Online Repository?</th>
<th>API?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content Standards</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDISC</td>
<td>• RESTful API</td>
<td>Metadata Repository</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>• Content standard downloads in XML, JSON, and RDF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• XML schemas and conformance rules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDI</td>
<td>• XML schema</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>• XML and RDF vocabularies</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Exchange Standards</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL7 FHIR®</td>
<td>• XML, JSON, and RDF schemas</td>
<td>N/A†</td>
<td>N/A†</td>
</tr>
<tr>
<td>HL7 CDA® R2</td>
<td>• XML schemas, schematron, and XSLT</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Terminology / Semantics Standards</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOINC®</td>
<td>• RELMA® application to search the LOINC database and map to LOINC codes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>• The LOINC table in Microsoft Access or CSV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• FHIR RESTful API pilot</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNOMED CT®</td>
<td>• Open source SNOMED CT online Browser</td>
<td>Terminology Browser</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>• SNOMED CT machine readable concept model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Mapping tool</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• RESTful API - not for healthcare system use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHO ICD</td>
<td>• Download ICD-10 in Classification Markup Language ICD-10 online browser</td>
<td>Terminology Browser</td>
<td>No</td>
</tr>
<tr>
<td>AMA CPT®</td>
<td>• Microsoft Access MDB and CSV download</td>
<td>Terminology Browser</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>• CPT® Developer's Tool Kit for download only</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• CPT® Assistant Online</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• SNOMED CT® to CPT® Rules-Based Cross Maps</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

† Schemas and API are fully documented and multiple reference implementations exist online.

* LOINC is piloting a FHIR-based RESTful API

** SNOMED CT provides a RESTful API for lookups, but it is not available for use in production healthcare systems

CDISC SHARE Exports. CDISC SHARE Exports make CDISC standards accessible via download on the CDISC Website\(^9\) in a variety of electronic formats including CSV, XML such as ODM-XML\(^{21}\) and Define-XML\(^{22}\), PDF and RDF. CDISC SHARE Exports help users find, understand and use rich metadata and controlled terminologies relevant to clinical studies more efficiently and consistently, and improve integration and traceability of clinical data from protocol through analysis. CDISC SHARE Export files are available to CDISC member organizations and to individual academic researchers upon request. Details on using CDISC SHARE Exports are available on the CDISC Website\(^{23}\).

Three categories of CDISC SHARE Export files exist. Normative content in one or more formats are made available as metadata. PDF guides, though not exclusively normative, are made available as documents provided to complement the machine-readable normative metadata and to make the standards artifacts available for download from one location on the CDISC Website. Diff files, or difference files, display differences between two versioned standards by listing the additions, deletions and updates from one version of a standard to the next. Diff files are intended to support impact analysis for adopters to assess what changes are present and how that would impact their systems. Diff files can be used to drive incremental loading of new standards versions into a standards MDR.

Discussion

Despite substantial annual investment in research, the time to bring new medical therapies to the bedside remains too long\(^{24}\). New approaches need to be taken to better leverage existing sources of health and research data\(^{25}\). Data from tens of thousands of interventional and non-interventional studies performed each year cannot be aggregated and analyzed without considerable cost, largely because the data are not collected using a standard format and semantics\(^{26}\).
Without data standards, commercial software tools that analyze the data cannot feasibly be produced requiring the development of bespoke software applications that contribute to higher costs, longer time-lines, and increased data quality concerns. Siloed, non-standard data may lead to redundant research, where previously-published discoveries not highlighted by Medline or search engines are likely to remain “hidden” in non-interoperable datasets that require informatics expertise to fully interpret\textsuperscript{27}. This redundancy and inefficiency ultimately prolongs the process of translation from bench to bedside to provide new treatments that may improve human health. Many scientists have also cited limited reproducibility as a major challenge to research utility\textsuperscript{28} that is exacerbated by limited access to study data that can be effectively used by others to verify study results\textsuperscript{29}. CDISC standards and other community-reviewed standards, ontologies, and code-sets provide tools that can help implementers solve these issues. Application of any standard requires an up-front investment on the part of an implementer, but once adopted these standards save resource time and money.

Due in part to regulatory requirements, the biopharmaceutical industry including major pharmaceutical companies, medical device companies, clinical research organizations, IT vendors, regulators, and others have implemented CDISC standards within their organization. Even before the December 2016 FDA mandate, over 70% of submissions to FDA were SDTM-compliant\textsuperscript{30}. CDISC standards implementation is not limited to regulated research sectors, which comprise only a segment of the total volume of clinical and translational research. Foundational and transport standards have been implemented within essentially all popular clinical research data systems, including the implementation of CDISC’s Operational Data Model (ODM-XML) in commercial applications\textsuperscript{31-34} and open source systems such as the OpenClinica\textsuperscript{35, 36} and REDCap\textsuperscript{37-39} electronic data capture platforms; within an importer to i2b2\textsuperscript{40}; and several others\textsuperscript{19}, CDISC SHARE will support more efficient incorporation of CDISC standards into tools such as these, helping to drive automation and innovation. CDISC’s SDTM and CDISC SHARE are being evaluated for their potential role in the to-be-developed NCI Clinical and Imaging Data Commons.

**Examples of CDISC SHARE’s Use.** In September 2016, CDISC collected survey data from over 200 CDISC SHARE users. This survey showed that over 50% of respondents used CDISC SHARE (1) to populate CDISC standards libraries in their own organizational MDRs, (2) to assess the normative content within standards, (3) to support impact analyses when considering adoption of a new version of a standard, and (4) to directly utilize the standards in software and reporting implementations.

**CDISC SHARE Pilots.** In 2015 and 2016, six organizations volunteered to participate in a CDISC SHARE API Pilot team that developed the first version of the API. The organizations included Business and Decision Life Sciences, EDETEK, Entimo, Accenture, Fujitsu, and the European Translational Information & Knowledge Management Services (eTRIKS\textsuperscript{42}), a European Union Innovative Medicines Initiative collaborative. All the pilot participants tested the CDISC SHARE API and many have since implemented interfaces into their commercial products. Many of the vendors participated in CDISC SHARE Showcases demonstrating software that loaded CDISC SHARE content via the API or using content from CDISC SHARE Exports. XML4Pharma, a CDISC SHARE Showcase participant, noted that updating their SDTM-ETL (Extract, Transform and Load) software with a new version of SDTM was reduced from 2 weeks to 2 hours using CDISC SHARE content.

**eTRIKS.** The Innovative Medicines Initiative (IMI)-funded European Translational Information and Knowledge Management Services (eTRIKS) project delivered an open, sustainable, standards-based translational research informatics and knowledge management platform and suite of tools and services for IMI-funded research and the greater community. eTRIKS utilized CDISC standards as part of its Standards Starter Pack and CDISC SHARE content within the metamodel of its Harmonization Service (eHS) Platform. This platform was created to aggregate and integrate clinical and biomedical data collected over the course of multiple, disparate protocols and projects. SDTM metadata were imported into the eTRIKS eHS to support data aggregation from multiple different studies.

**Trace-XML.** In order to support high quality data review and reproducibility in clinical research, it is critical to have access to an unbroken chain of data, ideally beginning at the electronic health record (EHR), that includes audit information\textsuperscript{43, 44}. The FDA has identified a lack of traceability as one of the top 7 data standards issues\textsuperscript{45}, yet essentially no solutions historically existed\textsuperscript{4}. Trace-XML has been developed to visualize traceability across the clinical research data lifecycle for an entire study\textsuperscript{46}. Using SHARE metadata, Trace-XML enables standardized clinical study metadata to be represented as an interactive graph displaying the full, interconnected history of each data element. Trace-XML makes full study traceability computable, and uses the CDISC SHARE API to add the additional metadata needed to instantiate the relationships between standards, such as linking a CDASH variable to the associated SDTM variable. Using the CDISC SHARE API enables Trace-XML to dynamically find and add the relationships between new CDISC content standards versions without requiring the software to maintain a local data store of that standards metadata.
**CDISC SHARE Ongoing Development.** The SHARE 1.0 model (Figure 2) provides a simplified view of CDISC standards and their components that is heavily abstracted. CDISC standards are interconnected to each other and to conformance rules, biomedical concepts, controlled terminology subsets, study activities, and other related metadata. To improve support for this additional metadata and to better document the relationships among them, the SHARE team has embarked upon a version 2.0 of the model that is anticipated for release in mid-2018. Examples of new features that will be available in CDISC SHARE 2.0 include (1) CDISC biomedical concepts (defined below), (2) machine-readable conformance rules, and (3) version 2.0 of the CDISC SHARE API. Requests from the community, including expanding SHARE to manage metadata from implementers and using the repository as a neutral mechanism by which to share protocol information, will be explored in 2018.

**CDISC Biomedical Concepts.** Normative CDISC standards content is generally represented as two-dimensional tables that render well in PDF and match reporting and aggregation needs, but do not always represent the way that organizations collect and store data. The MDR makes it possible for CDISC to expand those two-dimensional connections between variable and domain, for example, and to layer on multi-dimensional associations that reflect the way researchers and clinicians think about biomedical elements. These representations, created to model the rich complexity of translational research, are called biomedical concepts. A small number of more complex relationships are already available in CDISC SHARE for variables that bridge multiple domains, and a growing number of biomedical concepts exist in concept maps created using CMAP to represent relationships visually and template-based Excel files to represent relationships in a tabular format. Each biomedical concept includes a set of related concepts, defined relationships among concepts, subsets of controlled vocabularies, rules and logic. An example concept map from the CDISC Breast Cancer TA User Guide (Figure 3) highlights the unique observations and results that form the basis of associated biomedical concepts. Biomedical concept inputs may also be drawn from concept models developed for routine healthcare, such as the Clinical Information Modeling Initiative (CIMI).

![Figure 3. CDISC Breast Cancer TA Concept Map for Biomedical Concepts](image)

The CDISC SHARE 2.0 model has been extended to manage these biomedical concepts to provide sophisticated, machine-readable standards that will provide explicit value sets to populate drop-down lists in electronic case report forms, specify datatypes for a given concept, define which data must be collected together (e.g., a measurement and its unit of measure), and provide logic and validation rules.
**Machine-readable Conformance Rules.** CDISC Foundational and Transport Standards development teams have created conformance rules for their standards. The validation rules perform an important role in data quality within the CDISC standards as these rules often represent constraints specified or implied by the individual standards. The CDISC SHARE team will incorporate into SHARE 2.0 conformance rules required for different organizations to implement the standards consistently. These rules have become a necessary component for data exchange including regulatory submissions to the FDA and PMDA.

**Modern Transport Methods and Formats.** Regulators currently require the use of SAS version 5 transport files for submission, though this will eventually be deprecated in favor of more modern format. SHARE makes transport methodology or file format more immaterial, as CDISC SHARE provides the standards metadata in a range of modern electronic data formats. Via the CDISC SHARE API, software can be created by the community to consume the MDR’s content dynamically and present that content in whatever structure they choose. SHARE and ODMv2 provide API support that will enable more efficient and innovative ways to exchange clinical research data and metadata, including alignment with other advances in healthcare data exchange such as HL7’s FHIR.

**Conclusion**

The CDISC SHARE MDR provides normative CDISC standards metadata to increase the pace and quality of standards development, as well as to increase the efficiency and effectiveness of organizations implementing the standards. CDISC SHARE provides a novel model-based representation of the CDISC standards that has dramatically increased the number of relationships explicitly rendered within the normative standards and includes an end-to-end clinical research data lifecycle representation of the standards. The standards metadata in the CDISC SHARE model combined with the RESTful API to retrieve content enable the development of software tools that automate an increasing portion of the clinical research data lifecycle. New versions of the CDISC SHARE model and API will expand both the standards content as well as the services provided by the API to expand the opportunities available for software developers to create innovative metadata-driven applications.

**Glossary**

**Table 2.** Glossary of acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADaM</td>
<td>Analysis Dataset Model</td>
<td>JSON</td>
<td>Java Script Object Notation</td>
</tr>
<tr>
<td>AMA</td>
<td>American Medical Association</td>
<td>FDA</td>
<td>US Food and Drug Administration</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
<td>FHIR</td>
<td>Fast Healthcare Interoperability Resources</td>
</tr>
<tr>
<td>BRIDG</td>
<td>Biomedical Research Integrated Domain Group</td>
<td>MDB</td>
<td>Microsoft Database</td>
</tr>
<tr>
<td>CDA</td>
<td>Clinical Document Architecture</td>
<td>MDR</td>
<td>Metadata Repository</td>
</tr>
<tr>
<td>CDASH</td>
<td>Clinical Data Acquisition Standards Harmonization</td>
<td>NCI</td>
<td>National Cancer Institute (US)</td>
</tr>
<tr>
<td>CDISC</td>
<td>Clinical Data Interchange Standards Consortium (SDO)</td>
<td>NIH</td>
<td>National Institutes of Health (US)</td>
</tr>
<tr>
<td>CIMI</td>
<td>Clinical Information Modeling Initiative</td>
<td>ODM</td>
<td>Operational Data Model</td>
</tr>
<tr>
<td>CSV</td>
<td>Comma Separated Values</td>
<td>PMDA</td>
<td>Japan Pharmaceuticals and Medical Devices Agency</td>
</tr>
<tr>
<td>DDI</td>
<td>Data Document Initiative</td>
<td>PRM</td>
<td>CDISC Protocol Representation Model</td>
</tr>
<tr>
<td>Define-XML</td>
<td>CDISC standard for dataset metadata in XML format</td>
<td>RDF</td>
<td>Resource Description Format</td>
</tr>
<tr>
<td>EHR</td>
<td>Electronic Health Record</td>
<td>RELMA</td>
<td>Regensrif LOINC Mapping Assistant</td>
</tr>
<tr>
<td>EMA</td>
<td>European Medicines Agency</td>
<td>HL7</td>
<td>Health Level 7 (SDO)</td>
</tr>
<tr>
<td>EVS</td>
<td>Enterprise Vocabulary Services (NCI)</td>
<td>ICD</td>
<td>International Classification of Diseases</td>
</tr>
<tr>
<td>ISO</td>
<td>International Standards Organization</td>
<td>IMI</td>
<td>Innovative Medicines Initiative</td>
</tr>
</tbody>
</table>
Table 2 – Continued. Glossary of acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOINC</td>
<td>Logical Observation Identifiers Names and Codes</td>
<td>SNOMED CT</td>
<td>Systematized Nomenclature of Medicine - Clinical Terms</td>
</tr>
<tr>
<td>REST</td>
<td>REpresentational State Transfer</td>
<td>TA</td>
<td>Therapeutic Area</td>
</tr>
<tr>
<td>SDO</td>
<td>Standards Development Organization</td>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>SDTM</td>
<td>Study Data Tabulation Model</td>
<td>W3C</td>
<td>World Wide Web Consortium</td>
</tr>
<tr>
<td>SEND</td>
<td>Standard for Exchange of Nonclinical Data</td>
<td>XML</td>
<td>eXtensible Markup Language (W3C)</td>
</tr>
<tr>
<td>SHARE</td>
<td>Shared Health and Research Electronic library</td>
<td>XSLT</td>
<td>eXtensible Stylesheet Language Transformations</td>
</tr>
</tbody>
</table>

**Funding**

CDISC SHARE is partially funded through FDA U01 FC005854, U24 FD005243, and U24 FD005643; NIAID U01 FD005854 and from CDISC. Initial CDISC SHARE funding and ongoing support was provided by TransCelerate BioPharma. LB also received funding from the NCI Cancer Center Support Grant P30CA125123.

**References**

7. EMA-CTAG. Final advice to the European Medicines Agency from the clinical trial advisory group on clinical trial data formats. In: Group CTA, editor.: EMA; 2013.
41. Mendis M. CDISC ODM to i2b2 Importer Project2012. Available from: http://community.i2b2.org/wiki/display/ODM2i2b2/Home.
44. FDA. CDER common data standards issues document version 1.1. FDA; 2011.
From Sour Grapes to Low-Hanging Fruit: A Case Study Demonstrating a Practical Strategy for Natural Language Processing Portability

Stephen B. Johnson, Ph.D. 1, Prakash Adekkanattu, Ph.D. 2, Thomas R. Campion, Jr., Ph.D. 1,2, James Flory, M.D., M.S. 1, Jyotishman Pathak, Ph.D. 1, Olga V. Patterson, Ph.D. 3,4, Scott L. DuVall, Ph.D. 3,4, Vincent Major, M.S. 5, Yindalon Aphinyanaphongs, M.D., Ph.D. 5

1Healthcare Policy and Research, Weill Cornell Medicine, New York, New York; 2Information Technologies & Services, Weill Cornell Medicine, New York, New York; 3VA Salt Lake City Health Care System; 4University of Utah, Salt Lake City, UT; 5Center for Health Informatics and Bioinformatics, NYU Langone Medical Center, New York, New York

Abstract
Natural Language Processing (NLP) holds potential for patient care and clinical research, but a gap exists between promise and reality. While some studies have demonstrated portability of NLP systems across multiple sites, challenges remain. Strategies to mitigate these challenges can strive for complex NLP problems using advanced methods (hard-to-reach fruit), or focus on simple NLP problems using practical methods (low-hanging fruit). This paper investigates a practical strategy for NLP portability using extraction of left ventricular ejection fraction (LVEF) as a use case. We used a tool developed at the Department of Veterans Affair (VA) to extract the LVEF values from free-text echocardiograms in the MIMIC-III database. The approach showed an accuracy of 98.4%, sensitivity of 99.4%, a positive predictive value of 98.7%, and F-score of 99.0%. This experience, in which a simple NLP solution proved highly portable with excellent performance, illustrates the point that simple NLP applications may be easier to disseminate and adapt, and in the short term may prove more useful, than complex applications.

Introduction
Natural Language Processing (NLP) holds tremendous potential for patient care and clinical research 1-4. However, a recent review of the literature by Demner-Fushman and Elhadad suggests that NLP remains an “emerging technology”, with a significant gap between promise and reality 5. The NLP community has engaged in numerous challenge tasks in recent years, which have been beneficial in improving technical methods and research collaboration. But, due to the artificial nature of tasks suitable for such competitions, these efforts have had limited impact on real-world problems 6. Several studies have demonstrated success in portability of NLP technologies across institutions 7-10. However, a recent paper by Carrell et al. argues that there remain serious challenges in adapting NLP systems across multiple sites, which include assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions 11.

Carrell et al. suggest a variety of mitigation strategies, such as heuristic record linkage, acquisition of local knowledge, active learning, and tailoring with machine learning 11. In contrast, Demner-Fushman and Elhadad suggest sharing patterns for simple tasks and “more work on porting pipelines with easy domain adaptation” 5. These two strategies may be broadly contrasted as seeking hard-to-reach fruit (which may turn out to be sour grapes for some institutions) or low-hanging fruit, respectively. This paper investigates which factors might allow one to pursue the latter approach as a practical strategy for NLP portability.

Our project was motivated by the New York City Clinical Data Research Network (CDRN), a collaboration among six academic medical centers in the metropolitan area, seeking to collect and integrate clinical data to support patient-centered clinical research 12. The CDRN needed an approach that would leverage existing NLP resources at specific sites, while enabling sharing of resources across sites. The first main consideration was to select a system architecture for NLP based on standards, which has become a crucial strategy to facilitate portability and scalability 13-15. The second main consideration was to select a task with potential to be replicated across all sites. We chose left ventricular ejection fraction (LVEF), a primary diagnostic measurement of heart failure. LVEF is the ratio of the...
volume of blood ejected during systole to blood volume in the ventricle at the end of diastole. LVEF is typically measured by echocardiography and recorded in narrative text. A number of previous studies have shown success in extracting LVEF from clinical documents. 

Based on these factors, we chose to work with a system architecture called Leo, which was developed by the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI). Leo is a set of libraries that facilitate rapid development and scalable deployment of NLP systems, and builds upon the Apache Unstructured Information Management Architecture Asynchronous Scaleout (UIMA AS). In particular, this study focuses on a specific instance of Leo named Ejection Fraction Extractor (EFEx).

VINCI developed EFEx to extract LVEF values from clinical documents that originate at various centers within the VA. These studies were conducted entirely on VA documents, which raises a question about generalizability outside the VA system. However, with over 1,700 points of care and thousands of clinical authors, the VA system provides an exceptional data source for system training. Therefore, we expected that EFEx would be a better candidate for portability than a tool developed using data from a single medical center. This report details the initiative that we undertook to install and configure EFEx at Weill Cornell Medicine, and to extract LVEF from echocardiogram reports available in the MIMIC-III database.

Methods

Data source
We obtained echocardiograms from the Medical Information Mart for Intensive Care III (MIMIC-III) database. MIMIC is an openly available database developed by the MIT Lab for Computational Physiology. The latest version, MIMIC-III contains de-identified patient records for >40,000 critical care patients between 2001 and 2012. Researchers wishing to use the data must accept the data use agreement and provide evidence of completion of appropriate human subject research training. The MIT research team de-identified the data according to Health Insurance Portability and Accountability Act Privacy Rules, which included random date shifting, which preserves temporal relationships within a given patient but not across patients.

We extracted 8707 echocardiogram reports from the NOTEEVENTS table by selecting CATEGORY field for ‘Echo’. The table was filtered to only the first echocardiogram report, in chronological order, for each unique hospital admission (coded in MIMIC by HADM ID). In this study we restricted our data corpus to a single document type of echocardiograms (low-hanging fruit) and originated at an independent source, which in this case is Beth Israel Deaconess Medical Center. One consideration behind such a selection was to investigate the effectiveness of EFEx on documents originated at an independent source. This is important if we want to eventually deploy EFEx at other CDRN centers while maintaining same level of performance. MIMIC is an open source de-identified dataset not subjected to institutional review board approval. The performance of EFEx on echocardiograms originated within Weill Cornell Medicine is an ongoing study and will be the subject of a future report.

System description
Leo follows the model of UIMA AS with client and service components, along with an additional core library. The client defines inputs and outputs for processing and sends requests to the services. Setting up a client consists of selecting the required collection reader and listener, which could be a database or a local file system. The core contains tools that have been developed in conjunction with Leo to facilitate various NLP and annotation needs. The service component contains the server functionality for launching UIMA AS services. The service component also defines the type system and annotators as a pipeline architecture that implements all the logic necessary to extract a target information from unstructured documents. The basic architecture of Leo is shown in Figure 1, with the flow beginning at the reader.
Leo is built using the Java language and requires the Java runtime environment and the Apache package manager Maven, and can be installed on Windows, Linux, or Mac. We installed instances of EFEx running on Linux and Mac environments, and the setup procedure was essentially identical. The following steps were performed to create a fully functional EFEx instance.

1. We installed Java SDK 8 on our machines and set up an environment variable JAVA_HOME pointing to the JDK bin location, and added this to the PATH environment variable.
2. We installed Maven 3.3.9 and setup an environment variable MAVEN_HOME pointing to maven bin location, and added this to the PATH variable.
3. We downloaded UIMA-AS (http://uima.apache.org/downloads.cgi) and extracted the content to a suitable folder. We installed UIMA version 2.6.0 (uima-as-2.6.0-source-release.zip), and followed the instructions to compile and package the UIMA-AS. We set the UIMA_HOME environmental variable UIMA_HOME pointing to UIMA-AS root folder and added the bin folder to PATH variable.

The distribution package for EFEx was made available through a VA github repository. Installation of EFEx mainly involved downloading and extracting the content to a folder location on the machine. As part of the configuration setup, we created a folder called amq-broker under the uima-as folder and provided write permission to this folder. This folder is required for the broker service to copy all its configuration settings. The entire installation and basic configuration was completed in one day at WCM. However, the overall installation time may vary depending on technical skills available at individual centers.

Reference standard
The reference standard was developed at NYU Langone Medical Center. At WCM, all values were further confirmed through manual review of the entire document collection. Two reviewers examined each document on Excel spreadsheet. They were given training based on previously defined guidelines. These guidelines included identifying all mentions of LVEF and the associated quantitative values. If there were differences between the two reviewers’ findings, a third reviewer serving as adjudicator resolved the discrepancy. The reviewers also confirmed all documents that did not have any mention of LVEF information. We identified two values, EFmin and EFmax, corresponding to the lowest and the highest values of LVEF for each document in the dataset. The reviewer identified numeric values and ranges of LVEF (e.g. 55, 50-70), as well as severity-based descriptors such as normal, mild, moderate, and severe. The majority of reports had either a numerical value or a range of values. In documents that contain multiple instances of LVEF, we employed the following logic for determining the reference values:

1. A LVEF instance in the conclusion part, normally at the end of the document, takes precedence over one in the finding section.
2. A LVEF mention in the postoperative section takes precedence over one in finding or conclusion sections. (The postoperative section always follows the conclusion section in the document.)
In some reports, the LVEF value is expressed using a greater than or less than symbol (e.g. $LVEF > 55$). In these cases the reviewer extracted the value ignoring the symbol. Some echo reports express uncertainty about the LVEF value using a question mark (e.g. $LVEF ? 55-70$). In such cases, the reviewer extracted the value, provided that there was no other instance mentioned elsewhere in the document.

In reports where there was no quantitative value for LVEF available, we assigned a numerical value or a range of values using other information. LVEF concept synonyms were identified, including ‘lvef’, ‘left ventricular’, ‘LV’, ‘ejection fraction’, and modifiers were defined, such as ‘depressed’, ‘impaired’, ‘systolic dysfunction’ etc. When a concept was preceded or followed by modifiers to the severity level, such as mildly depressed, moderately depressed, or severely depressed, a quantitative value was assigned. Table 1 shows examples of modifiers and the corresponding values assigned. Despite using this mapping scheme, there were still documents with no concept-value pair identified in the reference standard. In general, these documents did not mention LVEF, or it was not possible to assign any meaningful value from the available information.

Table 1. Values assigned for concepts of ejection fraction with qualitative modifiers in developing the reference standard.

<table>
<thead>
<tr>
<th>Modifiers</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severely depressed</td>
<td>5-29</td>
</tr>
<tr>
<td>Moderately depressed</td>
<td>30-44</td>
</tr>
<tr>
<td>Mildly depressed</td>
<td>45-54</td>
</tr>
<tr>
<td>Grossly preserved</td>
<td>50-55</td>
</tr>
<tr>
<td>Normal</td>
<td>70</td>
</tr>
</tbody>
</table>

While developing the reference standard, the context as well as the overall content was taken into consideration in assigning a value or range of values to a concept. For example, there could be instances of LVEF expressed as quantitative values as well as qualitative descriptors, such as when the phrase ‘normal global systolic function’ was mentioned along with ‘severe regional left ventricular systolic function’ and ‘EF 20-25%’. In such cases, the numerical value took precedence over qualitative descriptors. Another example, when a document contained the phrases ‘moderately depressed LVEF’ and ‘(LVEF=30%)’ as well as ‘LVEF 70% previously, now 30%.’ In this case, the 30% is taken as the value for EFmin.

Extraction methodology
Patterson, et al. has described the logic for concept extraction employed in the present study in detail\(^2\). EFEx is a rule-based system that identifies the set of core concepts for LVEF using regular expressions, pattern matching, and filters. Because of the ambiguous nature of some of the concepts (such as ‘function’), the preceding text to each mention of the concept was used as a filter. Quantitative values were found using number patterns, but allowed for with or without modifiers such as ‘=’,” ‘(‘,’ ‘)’, ‘%,” ‘(<,’ and ranges of values.

Figure 2 shows the overall logic that was implemented in finding the concept-value pairs of LVEF. Steps A through K are used to extract concept-value pairs, if there is one found in the document. For those cases when no concept-value pair is identified through steps A to K, functionality was added to the original EFEx to look for qualitative modifiers used to describe the LVEF concept. This extended logic was implemented through steps M through O and effectively simulates the mapping scheme adopted in creating the reference standard. We identified 100 reports that were previously shown to have no output when processed by EFEx and used these as a training set for developing the extended logic. The output from the training set was manually reviewed to adjust the regular expression patterns through an iterative process.

Data analysis
We analyzed the current results on MIMIC-III data in two ways. In the first case, we analyzed data using the extraction logic implemented in the original EFEx for LVEF. This instance that was ported from VA has extraction logic implemented only through steps A to K as described in Figure 2. We refer this version as Original EFEx. Upon analyzing results on MIMIC data, we observed that the Original EFEx missed a significant number of documents where EF concept is described in a qualitative manner without any numerical value assigned. So at
WCM we further extended the extraction methodology by implementing additional logics to discover EF concept-value pair based on qualitative assessment through a mapping scheme. The extended logic implemented as steps M through O in Figure 2 improved the performance of EFEx significantly. The algorithm searched for both numeric values, ranges of LVEF (e.g. 55, 50-70) and severity-based descriptors based around the clinically relevant normal, mild, moderate, and severe labels. We refer this version as Extended EFEx. Performance measures were then calculated in both Original EFEx and Extended EFEx instances on the entire documents.

Figure 2. Extraction logic for LVEF implemented in EFEx system.

The results of EFEx output were tabulated and compared against the reference standard. Each document was classified as one of four possible cases: true positive (document had an LVEF mention and EFEx identified the concept-value pair and matched with the value given in the reference standard); false positive (document had no LVEF mention as given by a null value in reference standard, but EFEx produced a non-null concept-value pair); true negative (document had no LVEF concept-value mention as given by a null value in the reference standard, and EFEx did not find any concept-value pair); and false negative (document had an LVEF concept-value mention as given by a non-null value in the reference standard, but EFEx did not identify a concept-value pair, or the value extracted did not match with the corresponding reference standard value).

When there were multiple instances of concept-value pair extracted by EFEx, we used the following heuristic measures to select a given instance of LVEF in order to compare directly with the reference standard. We either selected the last one, normally in the conclusion part of the report (time usually moves forward in the report), or the lowest value (the disease typically worsens). The total outcomes of the four cases were then used to calculate various statistical performance measures. These included precision (positive predictive value), recall (sensitivity or
true positive rate), specificity (true negative rate), accuracy (number of correct identifications by the EFEx system divided by the number of documents the system analyzed), and the F-score (the harmonic mean of recall and precision).

Results

There were 8707 documents for analysis. Using the Original EFEx, we classified each document as one of four cases, for the purpose of calculating performance measures: true positive (6568), true negative (1124), false positive (0), and false negative (1015). These values resulted an overall accuracy of 88.3% (95% CI 87.7% - 88.9%), sensitivity of 86.6% (95% CI 85.8% - 87.4%), specificity of 100% (95% CI 99.6% - 100%), positive predictive value 100% (95% CI 99.9% - 100%), and an F-score of 92.8%. Percentage of severely and moderately depressed ejection fraction (LVEF < 45) cases is calculated to be 10.7%.

Using the Extended EFEx, we classified documents as true positive (7541), true negative (1026), false positive (98), and false negative (42). These values resulted an overall accuracy of 98.4% (95% CI 97.8% - 98.8%), sensitivity of 99.4% (95% CI 99.2% - 99.6%), specificity of 91.3% (95% CI 89.4% - 92.8%), positive predictive value 98.7% (95% CI 98.4% - 99.0%), and an F-score of 99.0. We observed an increased percentage of severely and moderately depressed cases for ejection fraction (18.1%).

Discussion

This experience illustrates how ejection fraction is an excellent example of ‘low hanging fruit’: a simple potential application for NLP that is relatively easily portable to new clinical settings. One limitation of this study is that the task of identifying LVEF measurements is relatively simple, with low variability of expressions and values to extract. In addition, the study examined only one document type. However, the relative simplicity of the task does not mean it is not important: ready availability of this important quantitative parameter has important implications for research, quality improvement, and clinical care.

The EFEx development team reported that the system achieved 98% positive predictive value and 93% sensitivity at the instance level across all medical centers across all VA. Garvin, et al., has developed an NLP system based on the UIMA architecture for extracting LVEF values from echocardiograms that are generated at four centers within the VA. They have reported for document-level classification of EF of <40% had a sensitivity of 98.4%, a specificity of 100%, a positive predictive value of 100%, and an F-score of 99.2%. Also system test results at a concept level it was reported a sensitivity of 88.9%, a positive predictive value of 95%, and an F-score of 91.9. It should be noted that the discovery logic that was developed in that study is not the one implemented in the present EFEx system, although they both share some common features. The present results on the MIMIC-III dataset show a comparable overall performance when analyzed with EFEx without the qualitative concept mapping. The results on the Extended EFEx showed an improved performance matching the document level values reported above.

Recently, Nath et al. has reported an NLP tool named EchoInfer for large-scale data extraction from echocardiography reports at a single medical center. They have reported a recall 95-99% and a precision > 96% for LVEF. When compared to the performance of EchoInfer the result obtained from the EFEx system shows a slightly lower performance, when using EFEx without any concept-mapping scheme in its discovery logic on MIMIC data. However, with the concept-mapping scheme, the performance of EFEx improved significantly and the values are slightly better than the EchoInfer reported above. For the entire dataset, the Original EFEx classified 1015 documents as false negative. However, with Extended EFEx, we observed only 42 false negative cases. Some of the documents where Extended EFEx failed to identify a value for LVEF are one in which both left and right ventricle is mentioned together in one statement. Similarly, for documents in which the target concept value immediately followed by a different numeral (e.g. a list index number), our extraction logic failed to identify the correct value for EF. The adoption of the mapping scheme significantly improved the identification of severely and moderately depressed cases in the dataset. With Extended EFEx, we observed a 40% increase in the number of cases with LVEF < 45, which were confirmed by the manual review. This substantial increase further supports the effectiveness of the extended logic that was implemented, as these additional cases would not have been discovered using the original logic alone.
On the flip side, implementation of the extended logic introduced several false positive cases for EF. While no false positive case was observed with Original EFEx, 98 false positive cases were observed on Extended EFEx. The mapping scheme we implemented does not assign values for cases such as mild to moderate depressed, moderate to severely depressed, borderline depressed, or more depressed. For documents with these statements, EFEx assigns incorrect values for EF. Similarly, a statement such as Preserved LVEF (effective forward LVEF may be depressed given the severity of valvular regurgitation) (HADM_ID = 182611) is subject to interpretation and no value is given in the reference standard. Our extended logic assigned a value of 45. Typical of most NLP systems, there is room for further improvement in the extraction logic as evident from some of the false positive and false negative cases observed with EFEx.

Recent papers have identified a number of challenges facing NLP portability\(^5\text{-}^{11}\), such as assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions. An additional challenge arises when using standardized NLP architectures such as UIMA, especially when integrating multiple NLP modules\(^{13}\text{-}^{15}\). A final challenge not identified by these papers involves leadership of innovators (“benchmark institutions”) to adopters rather than a pull from the adopter\(^{19}\).

The five challenges for NLP portability are summarized in Table 2, along with strategies for mitigating the challenges described in the cited literature. The strategies can be roughly partitioned into technologically advanced methods addressing complex NLP tasks (striving for the hard-to-reach fruit), and more practical methods addressing simpler NLP tasks (settling for the low-hanging fruit). Focusing on target concepts with low sensitivity to document location is found to be a good practical strategy for the portability of NLP tools. Our own experience showed that simple concepts where the associated values follow a general convention or prescribed format are good candidates for Leo. At WCM, our ongoing development effort resulted in other instances of Leo were we used this strategy effectively. We had achieved high performance in extracting PHQ-9 score from encounter notes. Similarly, we achieved high performance in extracting TNM stages, Gleason score and ICD-9/10 diagnosis codes from surgical pathology reports. In these cases, the precision and recall of Leo instances were sufficiently high enough, and we are currently in the process of making these data available in the i2b2 instance at WCM.

Table 2. NLP portability challenges, and mitigation strategies that require advanced methods (hard-to-reach), and more practical methods (low-hanging).

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Strategy: Hard-to-Reach</th>
<th>Strategy: Low-Hanging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assemble corpora with heterogeneous document types</td>
<td>Use heuristic linkage methods; develop document classifiers</td>
<td>Exploit metadata; focus on single document type</td>
</tr>
<tr>
<td>Navigate diverse report structures</td>
<td>Customize document segmentation algorithms; employ active learning</td>
<td>Select pattern with low sensitivity to document location</td>
</tr>
<tr>
<td>Analyze idiosyncratic linguistic expressions</td>
<td>Use machine learning to tailor complex patterns</td>
<td>Re-use or adapt simple patterns developed previously</td>
</tr>
<tr>
<td>Integrate multiple NLP modules</td>
<td>Employ large number of modules; adapt to meet architecture standards</td>
<td>Employ small number of modules; re-use use or adapt modules previously standardized</td>
</tr>
<tr>
<td>Lead the dissemination project</td>
<td>Acquire funding to support the innovator site; supply expertise in NLP methods</td>
<td>Draw on existing resources at the adopter site; use conventional software skills</td>
</tr>
</tbody>
</table>

**Conclusion**

We extracted LVEF information from echocardiogram reports from the MIMIC-III database using the EFEx NLP system. We compared the results to a reference standard developed manually by human reviewers. EFEx in its original version showed lower performance compared to the performance reported on VA documents that are different in document formats and content. However, when the extraction logic was modified to include a concept-
value mapping scheme similar to the mapping scheme used in developing the reference standard, EFEx had an accuracy of 98.4%, sensitivity of 99.4%, a positive predictive value of 98.7%, and an F-score of 99.0%. These values match reasonably well with that reported earlier on VA generated echocardiograms. The extended extraction logic also improved the discovery of cases having severely or moderately depressed LVEF by 40%. The current study on the LVEF extraction from the MIMIC dataset suggests that the EFEx performance varies depending on documents that are originated at different clinical settings.

The project described in this paper pursues a practical strategy to pursue a relatively simple NLP task (low-hanging fruit). We exploited database metadata to focus on single document type (cardiology reports). We chose a pattern with low sensitivity to document location (we used the last occurrence of LVEF). The adopter (WCM) led the dissemination project, drawing on existing resources, and employing conventional software skills. This case study provides evidence that an NLP system can be ported successfully from one institution to another, enable customization to a new data source, and achieve comparable performance. The identification of practical strategies for NLP portability has paved the way for sharing NLP tools among the multiple institutions in the NYC CDRN, and may provide useful guidance for other institutions interested in pursuing a similar approach.

References

11. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, Crockett SD, Gourevitch RA, Dean KM, Mehrotra A. Challenges in adapting existing clinical natural language processing systems to multiple,


19. Chung J, Murphy S. Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. AMIA Annu Symp Proc 2005:131e5


Predicting Causes of Data Quality Issues in a Clinical Data Research Network

Ritu Khare, PhD¹, Byron J. Ruth, MS¹, Matthew Miller, MS¹, Joshua Tucker¹, BS, Levon H. Utidjian, MD, MBI¹,², Hanieh Razzaghi, MPH¹, Nandan Patibandla, MS³, Evanette K. Burrows, BS¹, L. Charles Bailey, MD, PhD¹,²

¹Departments of Pediatrics and Biomedical & Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104; ²Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104; ³Information Services Department, Children’s Hospital Boston, Boston, MA, 02115

Abstract

Clinical data research networks (CDRNs) invest substantially in identifying and investigating data quality problems. While identification is largely automated, the investigation and resolution are carried out manually at individual institutions. In the PEDSnet CDRN, we found that only approximately 35% of the identified data quality issues are resolvable as they are caused by errors in the extract-transform-load (ETL) code. Nonetheless, with no prior knowledge of issue causes, partner institutions end up spending significant time investigating issues that represent either inherent data characteristics or false alarms. This work investigates whether the causes (ETL, Characteristic, or False alarm) can be predicted before spending time investigating issues. We trained a classifier on the metadata from 10,281 real-world data quality issues, and achieved a cause prediction F1-measure of up to 90%. While initially tested on PEDSnet, the proposed methodology is applicable to other CDRNs facing similar bottlenecks in handling data quality results.

Introduction and Background

Clinical data research networks (CDRNs) transform electronic health record (EHR) data from multiple institutions into common data models, and make that data available, either in a centralized or a distributed fashion, to conduct a wide range of scientific studies.¹-³ Given that EHRs are designed for clinical operations rather than research use, one of the most critical aspects in building a CDRN is to ensure that the aggregated clinical data are “high-quality” or “ready for research use.”⁴-⁷ CDRN datasets are typically built in an iterative fashion. The data coordinating center executes certain data characterization or validation modules on the dataset to identify any data quality problems; these problems are communicated to the contributing institutions that investigate and resolve the problems, and generate the improved datasets. The investigation of data quality problems is a complex process that involves local replication of the problems, reviews of the relevant extract-transform-load (ETL) code, verification of data assumptions, and discussions about local data characteristics with the interdisciplinary team of clinicians, analysts, researchers, and administrative staff.

In the past, several studies have recommended techniques for conducting data quality assessments on EHR-derived datasets, such as using expert judgment, heuristics, knowledge of impossibilities, gold standard benchmarking, code reviews, conformance to value set domains, and computation of derived values.⁵,⁸,⁹ and more recently Kahn et al. designed a comprehensive ontology to classify data quality checks.⁹ However, handling, analysis, or classification of real-world data quality problems or issues, are largely undocumented.¹⁰ Here, we present an automated approach that given a data quality issue, classifies the issue cause as “ETL” vs. “characteristic” vs. “false alarm,” to assist in prioritization and resolution of issues. The main contribution of this work is the use of supervised machine learning to predict the causes of data quality issues and achieve a promising performance.

In this study, we focus on a pediatric CDRN, PEDSnet, that aggregates EHR data from eight of the nation’s largest children’s hospitals¹¹,¹² using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM).¹³ PEDSnet has invested substantial efforts in designing and implementing data quality “checks” to evaluate the validity of EHR-derived datasets and identify any data quality “issues” that indicate that the data could be inaccurate or difficult to use for some research purposes.¹⁴ PEDSnet uses “GitHub issues” to report the data quality issues to individual sites as shown in Figure 1. Once the issues are reported, the originating site’s first task is to determine whether the issue is an error in the ETL programming pipeline (Figure 1a), or represents a characteristic or inherent property of data such as EHR data entry error, administrative issues, source data incompleteness, institutional data anomalies, etc. (Figure 1b), or is a false alarm caused due to a programming bug in the PEDSnet
The next task is to resolve any ETL-related issues for the next data submission. Based on a detailed analysis of the six most recent data cycles or iterations in PEDSnet, a partner site on an average investigates 26 new issues in each cycle, out of which only 35% represent resolvable problems (i.e. ETL category), as shown in Figure 2. In addition, based on the analysis of issue timelines in GitHub, we found that a data quality issue is open for 31 days on an average suggesting the potential duration of the issue investigation process; see Figure 3 for a distribution of GitHub issue duration across different types of causes. In sum, issue handling is a major bottleneck toward the iterative development of PEDSnet, as the partner sites need to resort to time-consuming and expensive processes for manual prioritization and investigation of individual issues. In this study, we examine whether the cause of a data quality issue can be predicted before delving into investigation, to help minimize issue fatigue and avoid spending time on issues that cannot be resolved, e.g. characteristic issues, or that should not have been reported at all, e.g. false alarms.

![GitHub screenshots of PEDSnet data quality issues illustrating different causes - top to bottom (a) ETL issue (in red), (b) Characteristic issue (in blue), (c) False alarm (in gray).](image)

Figure 1. GitHub screenshots of PEDSnet data quality issues illustrating different causes - top to bottom (a) ETL issue (in red), (b) Characteristic issue (in blue), (c) False alarm (in gray).
Methods

As our data source, we use the PEDSnet data quality issue warehouse. The warehouse contains metadata about data quality issues, and manually identified causes of those issues. The metadata includes the affected domain(s) and field(s) in the CDM, the tailored description of the issue, site information, the check type (Table 1) generating the issue, and the version of the CDM adopted for the given data cycle. It should be noted that the “characteristic” issues, once determined, get documented in the subsequent data cycles, but are not reported to the sites to avoid duplication of efforts.
Table 1. Some Examples of Data Quality Check Types and Issues in PEDSnet

<table>
<thead>
<tr>
<th>Check Type (alias: Name)</th>
<th>Example Data Quality Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>InconSource: Inconsistency with source</td>
<td>Distribution of NULL values in race_source_value does not match with the distribution of “No Information” concept in race_concept_id in Person table</td>
</tr>
<tr>
<td>InvalidValue: Value set violations</td>
<td>A non-standard concept used for populating the condition_concept_id field</td>
</tr>
<tr>
<td>UnexFact: Unexpected facts</td>
<td>A medication name entered into the location.zip field</td>
</tr>
<tr>
<td>ImplEvent: Implausible events</td>
<td>Found encounters with visit_start_date occurring after visit_end_date</td>
</tr>
<tr>
<td>CatOutlier: Categorical outlier</td>
<td>A patient with over 30,000 procedures</td>
</tr>
<tr>
<td>UnexTop: Unexpected most frequent values</td>
<td>“injection for contraceptive” as the most frequent procedure at a site</td>
</tr>
<tr>
<td>UnexDiff: Unexpected difference from the previous data cycle</td>
<td>Decrease in the number of deaths, or large increase (e.g. 2X) in the number of conditions</td>
</tr>
<tr>
<td>MissData: Missing data</td>
<td>Gestational age is not available for 70% of patients</td>
</tr>
<tr>
<td>MissFact: Missing expected facts</td>
<td>No “creatinine” lab record found in measurement table</td>
</tr>
</tbody>
</table>

We hypothesize that training a machine learning classifier, using meta-data about known issues, can help determine the cause of a data quality issue, and that the classifier can deliver performance sufficient to drive issue prioritization. As input, we selected a variety of features from the PEDSnet issue warehouse, intended to capture several aspects of an issue. Overall, 83 binary features were selected. The features types and instances are described below, and illustrated in Table 2.

- Domain: The CDM table where the issue was observed, e.g. Person, Care_Site, Location, Death, Condition_occurrence, Visit_payer, Visit_occurrence, Procedure_occurrence, Measurement, Drug_exposure, Measurement_organism, etc.
- Field Type: The type of field where the issue was observed, e.g. numerical fields, foreign keys, concept identifiers, source values, combination of fields, or others.
- Check Type: The type of data quality assessments conducted to identify the issue; some examples are shown in Table 1.
- Prevalence: The number of records affected by the issue, categorized as full (100%), high (30%-100%), medium (1%-30%), low (0%-1%), or unknown.
- Site: The site where the issue is observed, including one of the eight PEDSnet sites.
- CDM version upgrade: A boolean feature denoting whether the PEDSnet CDM version was upgraded since the previous data cycle.

Table 2. Features types and positive features for the example issues shown in Figure 1

<table>
<thead>
<tr>
<th>Feature Types</th>
<th>ETL issue (Fig. 1a)</th>
<th>Characteristic issue (Fig. 1b)</th>
<th>False alarm (Fig. 1c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain</td>
<td>Condition_occurrence</td>
<td>Visit_occurrence</td>
<td>Drug_exposure</td>
</tr>
<tr>
<td>Field Type</td>
<td>Concept identifier</td>
<td>Multiple</td>
<td>-</td>
</tr>
<tr>
<td>Check Type</td>
<td>UnexTop</td>
<td>ImplEvent</td>
<td>UnexDiff</td>
</tr>
<tr>
<td>Prevalence</td>
<td>Medium</td>
<td>Low</td>
<td>Medium</td>
</tr>
<tr>
<td>CDM version upgrade</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

We targeted two classification problems, binary (ETL vs. Non-ETL), and three-way (ETL vs. Characteristic vs. False alarm). We evaluated several classification methods including Naïve Bayes (NB), Decision tree (DT), Decision tree with boosting (DTB), k-Nearest neighbor (KNN), and support vector machine (SVM). We used Python implementation of the classifiers, and used the datasets extracted from the PEDSnet issue warehouse for training. Prior to choosing specific configurations for these learners, a “model grid search” was performed using the
GridSearchCV algorithm to find the optimal set of parameters for each of the target learners as shown in Table 3. The search was performed on 80% of the data using a five-fold stratified training set. Each combination was evaluated against a hold-out set to score the model.

**Table 3.** The learned parameters for various classifiers using a grid search

<table>
<thead>
<tr>
<th>Learner</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision tree + pruning (DT)</td>
<td>Max depth = 10</td>
</tr>
<tr>
<td>Decision tree + pruning + boosting (DT+B)</td>
<td>Max depth = 4 Estimators = 300</td>
</tr>
<tr>
<td>k-nearest neighbor (KNN)</td>
<td>K = 5 (binary), 3 (three-way)</td>
</tr>
<tr>
<td>Naïve Bayes (NB)</td>
<td>Class Priors = None</td>
</tr>
<tr>
<td>Support vector machine (SVM)</td>
<td>Kernel = linear Error term = 0.1 Tolerance = 0.001</td>
</tr>
</tbody>
</table>

**Results**

We used the July 2017 version of the PEDSnet issue warehouse\(^1\), which contains metadata on 11,434 data quality issues identified over a span of 30 months. We drew two datasets for experimentation, *all-issues-dataset* which includes 10,281 issues after filtering out the issues with unknown causes, and *unique-issues-dataset*, with 4,388 issues, that is a subset of *all-issues-dataset* prepared after excluding duplicate characteristic issues. The class label distributions across both datasets are: 14.37% (ETL), 81.78% (Characteristic), and 3.84% (False alarm); and 33.68% (ETL), 57.31% (Characteristic), and 9% (False alarm); respectively.

Figures 4 and 5 show the receiver operating characteristics (ROC) curves and performance measures for binary classification (ETL vs. Non-ETL), respectively. The results indicate that the k-nearest neighbor and decision tree with boosting algorithms could be promising choices for this problem. The classifiers trained on the *unique-issues-dataset* deliver higher F1 measure for the ETL issues, as compared to the *all-issues-dataset*, given the higher balancedness. Figure 6 shows the performance of classifiers on three-way classification problem. The performance for each class is higher than that of the binary classifiers. The classification performance on characteristic issues is higher than that on ETL or false alarms. This is most likely due to the availability of significantly higher training examples for characteristic issues in both the datasets.

![Figure 4. ROC curve for the binary (ETL vs. Non-ETL) classifier trained on all-issues-dataset](image-url)
Figure 5. Performance measures for binary (ETL vs. Non-ETL) classification of issues
Figure 6. Performance measures for three-way (Characteristic, ETL, False Alarm) classification of issues
To further understand the results, we examined the types of issues that constitute the most frequent error cases in the all-issues-dataset using the Decision tree with boosting classifier (Table 4). In general, the issues that are frequently difficult to classify tend to be limited to five major check types representing candidates for further study. In the majority of error cases, the classifier could not determine whether MissData (missing data) for drug_exposure and measurement was due to inherent characteristics or ETL error. Both these domains represent some of the most evolving domains in PEDSnet wherein the sites are gradually populating various fields, and hence the fluctuations in causes of missingness in the past several cycles. Another difficult check type was UnexDiff (unexpected difference in the number of records between two data cycles) wherein it is difficult to determine whether the issue is due to an ETL error or due to a natural enlargement of site’s dataset, i.e. false alarm.

### Table 4. Most frequent error cases (Check type, domain), FP=false positive, FN= false negative

<table>
<thead>
<tr>
<th>Cause Class</th>
<th>ETL</th>
<th>Characteristic</th>
<th>Non-issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UnexDiff, Measuremen</td>
<td>MissData, drug_expos</td>
<td>MissData, Measuremen</td>
</tr>
<tr>
<td>2</td>
<td>MissData, Measuremen</td>
<td>MissData, MissData, drug_expos</td>
<td>MissData, MissData, Measuremen</td>
</tr>
<tr>
<td>3</td>
<td>InvalidConID, Provider</td>
<td>MissConID, Drug_expos</td>
<td>MissConID, MissConID, Drug_expos</td>
</tr>
</tbody>
</table>

### Discussion

Based on our experience of conducting iterative data quality assessments on a pediatric CDRN, we find that a majority (>60%) of the data quality issues should receive lower priority for investigation, as they are either false alarms or an inherent characteristic of data that cannot be altered or resolved. In this study, we have studied the cause prediction problem using machine learning classifier that, given a data quality issue, predicts the cause of the issue. The best performing classifier achieved a promising F1-measure of 0.9, and indicates the potential to save significant effort by the data generation teams. While this study was primarily driven by the efficiency challenges faced in PEDSnet and the proposed method was tested on the pediatric dataset, the methodology can be applied to benefit other CDRNs.

By conducting the experiments using several classifiers with different class configurations, we were able to identify strong candidates for real-world implementation and execution. While our interest primarily lies in accurately predicting ETL issues, the performance on ETL issues (F1-measure, 0.71) of all classifiers left substantial scope for improvement, e.g. further analysis of the frequent use cases identified through error analysis. In the future, we plan to extend this work using systematic feature selection, development of more granular causal classes, development of balanced datasets, and assessment of the impact of automatic predictions on user experience.

### Acknowledgements

The authors would like to thank all the members of the PEDSnet data committee team, in particular, Daniel Eckrich, Janet Zahner, Kevin Matthews, Melody Kitzmiller, Richard Hoyt, Rob Brown, and Victoria Soucek, for participating in investigations of data quality issues and conducting discussions on GitHub. The authors are indebted to the patients and families contributing to the PEDSnet dataset. Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-01556).
References


Multi-Task Learning to Identify Outcome-Specific Risk Factors that Distinguish Individual Micro and Macrovascular Complications of Type 2 Diabetes

Era Kim, MS1,2, David S. Pieczkiewicz, PhD1, M. Regina Castro, MD3,
Pedro J. Caraballo, MD3, Gyorgy J. Simon, PhD1
1Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA;
2Medica Research Institute, Minnetonka, MN, USA and OptumLabs Visiting Fellow;
3Mayo Clinic, Rochester, MN, USA

Abstract

Because deterioration in overall metabolic health underlies multiple complications of Type 2 Diabetes Mellitus, a substantial overlap among risk factors for the complications exists, and this makes the outcomes difficult to distinguish. We hypothesized each risk factor had two roles: describing the extent of deteriorating overall metabolic health and signaling a particular complication the patient is progressing towards. We aimed to examine feasibility of our proposed methodology that separates these two roles, thereby, improving interpretation of predictions and helping prioritize which complication to target first. To separate these two roles, we built models for six complications utilizing Multi-Task Learning—a machine learning technique for modeling multiple related outcomes by exploiting their commonality—in 80% of EHR data (N=9,793) from a university hospital and validated them in remaining 20% of the data. Additionally, we externally validated the models in claims and EHR data from the OptumLabs™ Data Warehouse (N=72,720). Our methodology successfully separated the two roles, revealing distinguishing outcome-specific risk factors without compromising predictive performance. We believe that our methodology has a great potential to generate more understandable thus actionable clinical information to make a more accurate and timely prognosis for the patients.

Introduction

Type 2 Diabetes Mellitus (T2DM) is an irreversible chronic disease. It is associated with the metabolic syndrome, a cluster of interrelated conditions that include high blood pressure (BP), chronically elevated fasting plasma glucose (FPG), abdominal obesity, and lipids imbalance including elevated triglycerides (TG), and low high-density lipoprotein (HDL). Since complicated interactions among these conditions exist, even a minor adjustment on a single risk factor can dramatically influence the patient’s health status and clinical outcomes. Hence, a comprehensive understanding of the effects of these risk factors on various complications is necessary for the successful long-term management of T2DM patients.

Studies that identify risk factors for complications of T2DM abound, however they fail to paint an accurate picture of the patient’s health status and progression to the most likely next complication. In these studies, regardless of which complication they focus on, the risk factors tend to be largely the same (e.g., BP, FPG, lipids, and kidney function). The reason for this large overlap is that the above risk factors capture the effect of deteriorating overall metabolic health that underlies all these outcomes rather than capturing the effects that differentiate among the outcomes. This suggests that the risk factors have two roles: first, they describe the extent to which the patient’s overall metabolic health has deteriorated and second, they signal a particular complication that the patient is progressing to. Given that existing studies have focused on a single or occasionally a few complications and modeled them independently, they have not separated these two roles. Hence, it is difficult to know whether a risk factor is significant in progression to a particular complication or whether it merely describes the deterioration of overall metabolic health. To understand the direction of progression, namely, which of the many possible complications the patient is most likely to develop next, separating these two roles is critical.

The deterioration of underlying metabolic health is a commonality across all the complications. If we identify the commonality and remove it from the entirety of a risk factor’s effect, all that remains is outcome-specific effect. To model this, we had two challenges. First, to correctly capture the commonality, we needed to examine a wide range of complications using sufficient amounts of patient data. Because, if we study a single complication, the commonality is not identifiable so the distinction is lost. In this study, we used two independent datasets. As the primary, we had EHR data (N=9,793) collected from the University of Minnesota Medical Center (UMMC) and used them for model training and internal validation. As the secondary, we had claims and EHR data from the
OptumLabs Data Warehouse (OLDW) (N=72,720) and used them for external validation. Because these datasets contained years of medical history of a large number of patients, they offered sufficient amounts of patient data and allowed us to examine multiple complications simultaneously.

Second, it was methodologically challenging to isolate the commonality from the entirety of a risk factor’s effect because, it is not distinguishable. Multi-Task Learning (MTL) is a technique to model multiple related outcomes by exploiting their commonality\(^9,10\). In our case, modeling progression to each individual complication is a modeling task, and these tasks are related because deterioration in overall metabolic health underlies them all. We used MTL to integrate these tasks and identify the commonality among them. This approach is tantamount to applying MTL in reverse: rather than exploiting the commonality across the outcomes towards improved predictive performance, we discard the commonality to reveal differential markers, risk factors that are specific to each complication.

Considering that an accurate and timely prognosis for the patients often remains unsatisfactory\(^11\) and there is limited evidence available for clinical decision support, our methodology that improves the interpretation of predictions and generates more understandable clinical information will help prioritizing the outcomes and developing optimal individualized T2DM management.

**Materials and Methods**

*Primary Dataset for Training and Internal Validation*

We used 10-year, de-identified EHR data (Jan 1, 2004-Dec 31, 2013) including inpatient, outpatient, and emergency department visits from the University of Minnesota Medical Center (UMMC), a main university hospital, located in Minneapolis, MN. From the EHR data, we extracted patient demographics (age, gender), smoking status, vital signs (BP, pulse, and Body Mass Index BMI), lab results (HbA1c, lipid panel, Glomerular Filtration Rate GFR), three diagnoses comorbid to T2DM (dyslipidemia, hypertension, obesity), and six diagnoses of complications of interest: chronic kidney disease (CKD), acute renal failure (ARF), ischemic heart disease (IHD), congestive heart failure (CHF), peripheral vascular disease (PVD), and cerebrovascular disease (CVD). ARF is not usually associated with T2DM but involves organs or functions that are affected by T2DM. To demonstrate our proposed methodological validity, we intentionally included ARF as one of outcomes. We used 80% and 20% of UMMC data for model training and internal validation, respectively.

*Study Design and Cohort Selection*

We conducted retrospective cohort study. In UMMC data, we set up the study baseline at Jan. 1, 2010, collected patients’ 6-year medical history to create baseline patient characteristics, and followed them from baseline to Dec. 31, 2013, determining whether or not they developed any complication of interest (Figure 1).

![Figure 1. Study design](image)

None: No complication is developed. IHD: Ischemic Heart Disease. CKD: Chronic Kidney Disease. CVD: Cerebrovascular Disease. PVD: Peripheral Vascular Disease. CHF: Congestive Heart Failure. ARF: Acute Renal Failure.
Initially, we identified 22,946 adult T2DM patients based on ICD-9 codes. These patients were at least 18 years old at baseline, and they were generally diagnosed with T2DM within the 6-year period. When patients develop multiple complications, the effects of risk factors become conflated. Thus, we excluded 8,979 patients who already developed any of the complications before baseline and 914 patients who developed multiple complications during the follow-up period. This was because we wanted to start with simple data without such conflated effects and achieve our goal of examining the feasibility of our methodology able to separate the two roles. We excluded 1,152 patients who had no HbA1c measurements at all, 1,611 patients who had no BP, pulse, or BMI measurements at all, 494 patients who had no lipids information at all, and 3 patients without any known smoking status, resulting in 9,793 patients.

Second Dataset for External Validation

For external validation, we used claims and EHR data from the OptumLabs Data Warehouse (OLDW), which includes de-identified claims data for privately insured and Medicare Advantage enrollees in a large, private, U.S. health plan, as well as de-identified EHR data from a nationwide network of provider groups. The database contains longitudinal health information on enrollees, representing a diverse mixture of ages, ethnicities and geographical regions across the United States. The health plan provides comprehensive full insurance coverage for physician, hospital, and prescription drug services. The EHR data sourced from provider groups reflects all payers, including uninsured patients. We extracted 10-year data (Jan 1, 2006-Dec 31, 2015) from the OLDW and identified 72,720 T2DM patients using the same study design (Figure 1) and selection procedure.

Baseline Patient Characteristics in UMMC and OLDW Datasets

Table 1 shows baseline patient characteristics in UMMC and OLDW datasets with variables available in this study. These variables represent risk factors and are used henceforth. UMMC patients had similar HbA1c but higher SBP and DBP compared to US adults with diabetes (HbA1c, SBP, and DBP are 7.2%, 131.5mmHg, and 69.4mmHg, respectively), and they had signs of established CKD based on GFR. Compared to UMMC patients, OLDW patients were older and had better HbA1c, better lipids, better kidney function, but higher SBP and DBP.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>UMMC (N=9,793)</th>
<th>OLDW (N=72,720)</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>Male</td>
<td>51</td>
<td>46</td>
</tr>
<tr>
<td>age</td>
<td>Age (years)</td>
<td>58±13</td>
<td>60±12</td>
</tr>
<tr>
<td>never_smoker</td>
<td>Non-smoker</td>
<td>56</td>
<td>45</td>
</tr>
<tr>
<td>a1c</td>
<td>HbA1C</td>
<td>7.2±1</td>
<td>7.0±1</td>
</tr>
<tr>
<td>ldl</td>
<td>LDL-cholesterol (mg/dL)</td>
<td>103±28</td>
<td>101±28</td>
</tr>
<tr>
<td>hdl</td>
<td>HDL-cholesterol (mg/dL)</td>
<td>44±12</td>
<td>46±12</td>
</tr>
<tr>
<td>trigl</td>
<td>Triglycerides (mg/dL)</td>
<td>172±90</td>
<td>169±17</td>
</tr>
<tr>
<td>tchol</td>
<td>Total-cholesterol (mg/dL)</td>
<td>181±34</td>
<td>179±34</td>
</tr>
<tr>
<td>gfr</td>
<td>Glomerular Filtration Rate</td>
<td>58±32</td>
<td>76±27</td>
</tr>
<tr>
<td>gfr_norm</td>
<td>Normal Glomerular Filtration Rate</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>bmi</td>
<td>Body Mass Index (kg/m²)</td>
<td>34±7</td>
<td>34±8</td>
</tr>
<tr>
<td>sbp</td>
<td>Systolic Blood Pressure (mmHg)</td>
<td>127±11</td>
<td>131±11</td>
</tr>
<tr>
<td>dbp</td>
<td>Diastolic Blood Pressure (mmHg)</td>
<td>75±7</td>
<td>77±7</td>
</tr>
<tr>
<td>pls</td>
<td>Pulse (bpm)</td>
<td>76±9</td>
<td>77±9</td>
</tr>
<tr>
<td>hyperlip</td>
<td>Hyperlipidemia</td>
<td>81</td>
<td>86</td>
</tr>
<tr>
<td>htn</td>
<td>Hypertension</td>
<td>71</td>
<td>81</td>
</tr>
<tr>
<td>obese</td>
<td>Obesity (BMI &gt; 30)</td>
<td>70</td>
<td>67</td>
</tr>
</tbody>
</table>

Developing Multi-Task Learning Methodology

Under our hypothesis, each risk factor played two roles. The first role quantified the extent to which the patient’s metabolic health had deteriorated, and the second role signaled which complication the patient was most likely to develop next. The first role was common across all complications (common effect), and the second role was specific to each complication (outcome-specific effect).
Formally, given a design matrix $X$ that contained patients as rows and variables as columns, and $t$ measuring time to event (complication or censoring), we simultaneously built the following six models, one for each complication $c$

$$D^c: \quad \lambda^c(t) = \lambda_0^c(t) \exp(Xa) \exp(X\beta^c)$$

subject to $\|a\|_1 \leq C_1$ and $\|\beta^c\|_1 \leq C_2$

where $\lambda^c(t)$ was the patient’s hazard of developing complication $c$ at time $t$, $\lambda_0^c(t)$ was a complication-specific baseline hazard, $C_1$ and $C_2$ were user-defined thresholds, chosen via cross validation, and $\|\cdot\|_1$ denoted the L-1 norm (LASSO-penalty). $X$ contained all variables in Table 1 except T2DM-comorbidities (hyperlipidemia, hypertension and obesity) since their defining factors (lab results and vital signs) were included.

Conceptually, each $D^c$ model could be separated into two submodels as

$$\lambda^c(t) = \{k_0(t) \exp(X\alpha)\} \{k^c_1(t) \exp(X\beta^c)\}$$

where the first submodel (with coefficients $\alpha$) was a Cox model capturing the common effects, and the second submodel (with coefficients $\beta^c$) was a Cox model capturing outcome-specific effects for each complication $c$. We called the first submodel General Progression Model and the second submodel Differential Progression Model.

Since these two models used the same set of variables, coefficients $\alpha$ and $\beta^c$ were generally not identifiable (for each complication $c$, the effects of $\alpha$ and $\beta^c$ were not distinguishable). While the first of the two LASSO constraints in Eq. (1) simply induced sparsity in General Progression Model with the purpose of performing variables selection, the second LASSO-penalty made Differential Progression Models identifiable as it shrunk $\beta^c$ coefficients towards 0 forcing General Progression Model to explain as much of the variability as possible. We iteratively updated $\alpha$ and $\beta^c$ coefficients until they were stabilized (squared differences of coefficients between previous and current iterations were effectively zero).

If the entirety of a variable’s effect was only general deterioration of the metabolic health, its $\beta^c$ would be exactly 0. Conversely, if $\beta^c > 0$, the variable increased the risk of complication $c$ by $\beta^c$ from $\alpha$ (harmful); and if $\beta^c < 0$, it decreased the risk of complication $c$ by $\beta^c$ from $\alpha$ (protective). Therefore, non-zero $\beta^c$ coefficients identified differential markers; these were the risk factors that had effects beyond General Progression and enabled improved interpretation of progression to the most likely next complication.

**Internal and External Validation**

To determine the significance of $\alpha$ and $\beta^c$ coefficients, we performed 1,000 permutation tests and calculated empirical p-values. The key idea of permutation test was that variables were independent of randomly permutated labels; thus, coefficients of permutated labels were expected to have weaker associations than those of true labels. Then, the p-value of a coefficient could be calculated as the ratio of the number of permutation tests resulting in a stronger association to the total number of permutation tests. We internally evaluated predictive performance of our models in 20% of UMMC data and externally evaluated it in OLDW data using concordance index (c-index), typically used to assess predictive performance of Cox models. In internal validation, we also performed 1,000 bootstrapping with sample size of 100% UMMC patients to obtain 95% confidence intervals (95CIs). To demonstrate that we did not suffer a loss of performance due to our proposed MTL-based methodology, we compared predictive performance between ours and a reference methodology that built six independent models (LASSO-penalized Cox regression) for the six complications at a time.

**Results**

In this section, we are presenting results from our proposed methodology focusing on improved interpretation of risk factors and predictive performance in comparison with the reference methodology.

**Coefficients from Multi-Task Learning Methodology**

Figure 2 presents $\alpha$ and $\beta^c$ coefficients from General Progression and Differential Progression Models. The rows are the variables. The first column corresponds to $\alpha$ coefficients from General Progression Model and the remaining columns correspond to $\beta^c$ coefficients from Differential Models for each complication $c$. The interpretation of the coefficients is analogous to the regular Cox models: the exponent of a coefficient is the hazard ratio (HR) that the variable confers on the patient.
For most variables (e.g., HbA1c, LDL) which higher values are associated with higher risks, if \( \alpha > 0 \), it indicates a harmful association; and if \( \alpha < 0 \), it indicates a protective association with General Progression. In Differential Progression, if \( \beta^c > 0 \), the variable increases the risk of complication \( c \) by \( \beta^c \) from \( \alpha \), making the variable more important (harmful effect becomes larger); and if \( \beta^c < 0 \), the variable decreases the risk of complication \( c \) by \( \beta^c \) from \( \alpha \), making the variable less important (harmful effect becomes smaller). There are also variables (HDL, GFR, normal GFR, and never smoker) which higher values are associated with lower risks; thus, the interpretation of their coefficients is opposite. For example, if \( \alpha \) of GFR > 0, it means that higher GFR is protective of General Progression; if \( \beta^c \) of GFR > 0, it indicates that higher GFR is more important in progression to complication \( c \) (prospective effect becomes larger).

To help detecting significant \( \alpha \) and \( \beta^c \) coefficients, we visualized associations between variables and complications (harmful, protective, more important, and less important) and p-values (Figure 2). Coefficients with a circle are statistically significant, and those without are insignificant. Larger circles indicate smaller p-values (more significant). The exact p-values can be found in appendix Table A-1.

As expected, most variables significantly predicted General Progression: HbA1c, triglycerides, total cholesterol, GFR, normal GFR, BMI, SBP, DBP, non-smoker, and age (Figure 2). Traditionally, higher DBP is known to be harmful. But, several recent studies showed that DBP was protective of cardiovascular disease especially for older adults\(^6\). We also found that DBP is protective of General Progression. General Progression Model was a latent model in the sense that it did not have an observable outcome; it described the extent of deterioration in overall metabolic health. These variables of General Progression were those that many studies found to be significantly associated with an increased risk of micro and macrovascular complications and all-cause mortality\(^17,18\).

What is General Progression Model?

We defined General Progression mathematically as the effects that were common across all the complications and explained that General Progression captured deteriorating overall metabolic health. As an alternative, we interpreted \( \alpha \) coefficients as the log HR of progression to any complication. To illustrate this, we built a LASSO-penalized Cox model that predicted the development of any complication (this model had an event if a patient developed any complication) and compared coefficients from this model (Figure 3) with \( \alpha \) coefficients from General Progression Model (Figure 2). We found that they were similar with respect to effect size, sign, and significance, and this suggested that General Progression could be indicative of progression to any complication.

---

**Figure 2. Coefficients from Multi-Task Learning Methodology**

<table>
<thead>
<tr>
<th>Variable</th>
<th>General</th>
<th>CKD</th>
<th>ARF</th>
<th>IHD</th>
<th>PVD</th>
<th>CHF</th>
<th>CVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>0.0385</td>
<td>0.0407</td>
<td>0.0000</td>
<td>0.0644</td>
<td>0.0126</td>
<td>0.0088</td>
<td>0.0000</td>
</tr>
<tr>
<td>LDL</td>
<td>0.0000</td>
<td>-0.0003</td>
<td>0.0000</td>
<td>0.0034</td>
<td>0.0090</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>HDL</td>
<td>-0.0023</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-0.0010</td>
<td>0.0092</td>
<td>0.0027</td>
<td>0.0000</td>
</tr>
<tr>
<td>trigl</td>
<td>0.0007</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0003</td>
<td>0.0032</td>
<td>0.0000</td>
</tr>
<tr>
<td>LogHDL</td>
<td>-0.0020</td>
<td>-0.0014</td>
<td>0.0000</td>
<td>-0.0026</td>
<td>0.0092</td>
<td>0.0018</td>
<td>0.0000</td>
</tr>
<tr>
<td>LogGFR</td>
<td>0.0291</td>
<td>-0.0385</td>
<td>0.0000</td>
<td>0.0421</td>
<td>0.0614</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>GFR_norm</td>
<td>-2.3385</td>
<td>-3.4771</td>
<td>0.0000</td>
<td>3.5403</td>
<td>4.5379</td>
<td>0.0185</td>
<td>0.0000</td>
</tr>
<tr>
<td>bmi</td>
<td>0.0036</td>
<td>0.0050</td>
<td>0.0100</td>
<td>-0.0076</td>
<td>-0.0076</td>
<td>0.0463</td>
<td>0.0000</td>
</tr>
<tr>
<td>pbs</td>
<td>0.0092</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-0.0024</td>
<td>0.0037</td>
<td>0.0324</td>
<td>0.0000</td>
</tr>
<tr>
<td>sbp</td>
<td>0.0044</td>
<td>0.0033</td>
<td>0.0000</td>
<td>-0.0151</td>
<td>0.0146</td>
<td>0.0256</td>
<td>0.0000</td>
</tr>
<tr>
<td>dbp</td>
<td>-0.0003</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-0.0113</td>
<td>-0.0515</td>
<td>0.0020</td>
<td>0.0000</td>
</tr>
<tr>
<td>never_smoker</td>
<td>-0.1401</td>
<td>0.0626</td>
<td>-0.1632</td>
<td>-0.0482</td>
<td>-0.3956</td>
<td>-0.3487</td>
<td>0.0000</td>
</tr>
<tr>
<td>age</td>
<td>0.0137</td>
<td>0.0000</td>
<td>0.0000</td>
<td>-0.0043</td>
<td>0.0498</td>
<td>0.0510</td>
<td>0.0180</td>
</tr>
<tr>
<td>male</td>
<td>0.0292</td>
<td>0.0004</td>
<td>0.0000</td>
<td>-0.1407</td>
<td>0.5426</td>
<td>-0.2188</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
After achieving the overarching goal of our proposed methodology to separate common effects ($\alpha$) and outcome-specific effects ($\beta^c$) of a risk factor, we examined if results and their interpretations from our models clinically made sense. Especially, we wanted to have some of them consistent with known facts because if they were not, the utility of our methodology could be in doubt. To demonstrate this, let us consider the role of HbA1c in progression to CKD and IHD as an example as it is commonly accepted facts in practice: hyperglycemia is a key driver of microvascular complications (e.g., CKD), while dyslipidemia is a key driver of macrovascular complications (e.g., IHD)\textsuperscript{10}. General Progression showed that a unit increase in HbA1c conferred a HR of 1.039 (exp(.0385)) on all complications uniformly (Figure 2, row1, column1). However, higher levels of HbA1c ultimately affect the different complications differently. As mentioned, it is well-known that HbA1c is more predictive of CKD than IHD. Indeed, Differential Progression for CKD showed that a unit increase in HbA1c conferred an additional log HR of .0407 on patients, increasing the HR of CKD from 1.039 to 1.0824 (exp(.0385+.0407) (Figure 2, row1, column2).

It is also known that HbA1c is not as important in IHD as in CKD. Differential Progression for IHD showed that patients with higher HbA1c tended to suffer other (microvascular) complications. The log HR of IHD that a unit increase in HbA1c conferred on patients was negative, which decreased the HR of IHD from 1.039 to 0.9842 (exp(.0385-.0544)) (Figure 2, row1, column4). What it means is that patients with higher HbA1c are more likely to progress to a complication than patients with lower HbA1c, and that complication is less likely to be IHD but more likely to be a microvascular complication such as CKD. That is, General Progression described the patient’s tendency to progress to a complication, and the Differential Progression helped to target which complication the patient is more likely to develop next.

Differential Markers of CKD, IHD, PVD and CHF

To easily detect distinguishing patterns of differential markers, we visualized each of CKD, IHD, PVD and CHF as a series of spider plots\textsuperscript{12} in Figure 4. In a spider plot, variables are arranged as axes extending radially from a central point, and each observation makes a closed polygon connecting points on all of the axes. Emphasis is upon discerning the characteristic shapes of these polygons among observations, rather than extracting specific values. The interpretation of our plots is as follows. Each plot corresponds to a complication. Ten variables for vital signs and lab results construct individual axes, radially arranged around a center point. The $\beta^c$ coefficient of each variable is depicted by an anchor (node) on an axis. As higher values of HDL, GFR, and DBP are protective, the sign of coefficients of them are reversed only for visualization purposes. The same color encoding was used to identify significantly more or less important differential markers. For each variable, distance from the center indicates an increased risk. In each plot, a navy line connecting the $\beta^c$ coefficients represents Differential Progression, while a green line connecting zero on each axis conceptually represents General Progression, a reference for Differential Progression. By comparing these two lines on each axis, differential risk for complication $c$ beyond or below General Progression is easily distinguished. As we focused on straightforward interpretation, we did not perform normalization; thus, the scales of the variables are not comparable with each other. What is important is whether the navy line (Differential Progression) is outside or inside the green line (General Progression) on each axis.
IHD, PVD, and CHF are well-known concomitant macrovascular complications. They share similar pathophysiology and are believed to have similar risk factors. Given these facts, distinguishing among them without a methodology like ours is more difficult. In Figure 4, spider plots show distinguishing patterns of differential markers among these very similar diseases. In progression to IHD, LDL was more important; SBP, and lower DBP were less important. In progression to PVD, SBP and lower DBP were more important; BMI was less important. In progression to CHF, lipid abnormalities were less important; BMI, pulse, and SBP were more important (irregular or fast pulse is one of the symptoms of CHF).

**Coefficients from Reference Methodology**

Figure 5 shows coefficients from reference models. The rows are the variables, and the columns are complications. If a coefficient > 0, it indicates a harmful association; and if a coefficient < 0, it indicates a protective association with a complication c. In reference models, the two roles of a variable (α and βc coefficients) were not identifiable; thus, only the entirely of a variable’s effect was estimated, and the outcome-specific effect was masked. This was the motivation of our study and the key difference from our proposed methodology. To help detecting significant coefficients, we visualized the association between variables and complications (harmful and protective) and p-values. The exact p-values can be found in appendix Table A-2.

**Utility of Our Proposed Multi-Task-Learning Methodology**

To demonstrate clinical utility of our methodology in comparison with reference methodology, let us take ARF as an example. Although a major cause of ARF is not diabetes, reference models identified virtually all the variables to be predictive of ARF (Figure 5). While, Differential Progression Model for ARF showed that progression to ARF was only associated with the underlying advanced metabolic deterioration (General Progression), and all the variables were not specific to ARF (Figure 2).

Another example is CKD. Risk factors of CKD are well-understood. The reference model for CKD identified HbA1c (barely), GFR and age as significant risk factors, and they are indeed known risk factors. In fact, reference models identified age as a risk a factor for every complication; however, it is not that a patient is more likely to develop CKD just because he is older. Whereas, General Progression Model and Differential Progression Model for
CKD suggested that older patients were more likely to have their metabolic health deteriorated than younger patients; and, age played no role in progression to CKD beyond General Progression.

**Internal and External Validation**

Table 2 presents predictive performance in C-Index of our MTL-based models and reference models. Generally, they achieved similar predictive performance. Minimal albeit statistically significant differences were only observed in complications with small number of progressing patients.

For both, predictive performance was lower in external validation. UMMC data consisted of smaller number of patients from one healthcare system. Thus, they might be less representative of T2DM population than OLDW patients, or they might be subpopulation of OLDW patients. Also, patient characteristics differed fundamentally between them (Table 1). However, except CKD, C-Indices were still within 95CIs.

Table 2. Predictive Performance in C-Index (95CIs)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Methodology</th>
<th>CKD</th>
<th>ARF</th>
<th>IHD</th>
<th>PVD</th>
<th>CHF</th>
<th>CVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
<td>MTL</td>
<td>.74(.73-.79)</td>
<td>.58(.48-.82)</td>
<td>.57(.52-.58)</td>
<td>.75(.59-.80)</td>
<td>.83(.70-.91)</td>
<td>.75(.63-.78)</td>
</tr>
<tr>
<td>(UMMC)</td>
<td>Reference</td>
<td>.74(.73-.79)</td>
<td>.62(.48-.79)</td>
<td>.57(.52-.58)</td>
<td>.75(.60-.81)</td>
<td>.84(.67-.91)</td>
<td>.78(.65-.80)</td>
</tr>
<tr>
<td>External</td>
<td>MTL</td>
<td>.71</td>
<td>.61</td>
<td>.53</td>
<td>.61</td>
<td>.73</td>
<td>.64</td>
</tr>
<tr>
<td>(OLDW)</td>
<td>Reference</td>
<td>.71</td>
<td>.63</td>
<td>.53</td>
<td>.61</td>
<td>.74</td>
<td>.68</td>
</tr>
</tbody>
</table>

**Discussion**

Given that the effect of deteriorating overall metabolic health is common across all the complications, we hypothesized each risk factor had two roles: describing the extent of deteriorating overall metabolic health and signaling a particular complication the patient is progressing towards. We have successfully demonstrated that our proposed methodology separated these two roles of risk factors and revealed distinguishing patterns of differential markers. Also, we modeled multiple complications simultaneously by sharing their information; thereby, generating systematic and comprehensive interpretation of different roles of risk factors among various complications.

Our study has important strengths. First, we made more understandable predictions for clinicians by focusing on improved interpretability. Usually, high predictive accuracy is of key importance in prediction models. However, lack of clarity in the interpretation of predictions limits their usefulness in practice. Second, we externally evaluated predictive performance of our models, which were rarely done in other studies. Although the reference model did well or slightly better than our proposed model, the difference was minimal. So, we would say that we did not compromise predictive performance due to our proposed methodology. Third, our methodology is of high utility because it can be applied to other clinical conditions in which comorbidities matter.

We have several limitations. When identifying cohorts, we excluded patients who developed multiple complications. Although this action limits the generalizability of our work, our primary interest was to demonstrate the feasibility and utility of our proposed methodology. Additionally, we excluded patients with unknown vital signs, lab results, and/or smoking status. We tested differences between final study cohort and these excluded patients. All variables except hdl, tchol, dbp, never_smoker, and age were significantly different, and the excluded patients were generally sicker. Thus, our study is subject to selection bias. But, ours was not to estimate the effect of a risk factor, in which addressing selection bias using imputation methods is critical, but to separate the entirety of the effect into common and outcome-specific effects. Lastly, we used variables easily obtained from EHR data. Many studies have found that T2DM is disproportionally affected by race, ethnicity and/or socioeconomic status. Although they are very important risk factors, we mainly focused on modifiable risk factors.

When we build a model on large amounts of data, most variables become statistically significant; however, they may be clinically irrelevant. To impact individualized patient care, it is critical to develop enabling technologies that extract clinically useful information from the large amounts of data. Our future work is to extend this work to larger cohorts and overcome the limitations. If we can obtain reasonable levels of generalizability, we believe that our methodology will have significant potential to help clinicians prioritizing outcomes and making a more accurate prognosis for T2DM patients.
Acknowledgements

This work was partially supported by the NIH award LM011972 and the NSF award IIS 1602198. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References


### Appendices

<table>
<thead>
<tr>
<th>Variable</th>
<th>General</th>
<th>CKD</th>
<th>ARF</th>
<th>IHD</th>
<th>PVD</th>
<th>CHF</th>
<th>CVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1c</td>
<td>0.008</td>
<td>0.049</td>
<td>0.340</td>
<td>0.027</td>
<td>0.005</td>
<td>0.058</td>
<td>0.327</td>
</tr>
<tr>
<td>fdl</td>
<td>0.329</td>
<td>0.117</td>
<td>0.289</td>
<td>0.034</td>
<td>0.265</td>
<td>0.250</td>
<td>0.256</td>
</tr>
<tr>
<td>hdl</td>
<td>0.058</td>
<td>0.281</td>
<td>0.343</td>
<td>0.123</td>
<td>0.145</td>
<td>&lt; 0.001</td>
<td>0.283</td>
</tr>
<tr>
<td>trigl</td>
<td>0.014</td>
<td>0.297</td>
<td>0.397</td>
<td>0.059</td>
<td>0.079</td>
<td>&lt; 0.001</td>
<td>0.279</td>
</tr>
<tr>
<td>tchol</td>
<td>0.021</td>
<td>0.057</td>
<td>0.25</td>
<td>0.039</td>
<td>0.023</td>
<td>0.059</td>
<td>0.234</td>
</tr>
<tr>
<td>gfr</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.265</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.251</td>
<td>0.255</td>
</tr>
<tr>
<td>gfr_norm</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.261</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.251</td>
<td>0.275</td>
</tr>
<tr>
<td>bmi</td>
<td>0.042</td>
<td>0.067</td>
<td>0.333</td>
<td>0.051</td>
<td>0.034</td>
<td>&lt; 0.001</td>
<td>0.272</td>
</tr>
<tr>
<td>pls</td>
<td>0.188</td>
<td>0.290</td>
<td>0.320</td>
<td>0.108</td>
<td>0.076</td>
<td>&lt; 0.001</td>
<td>0.294</td>
</tr>
<tr>
<td>sbp</td>
<td>0.01</td>
<td>0.076</td>
<td>0.31</td>
<td>0.002</td>
<td>0.083</td>
<td>&lt; 0.001</td>
<td>0.272</td>
</tr>
<tr>
<td>dbp</td>
<td>0.006</td>
<td>0.261</td>
<td>0.307</td>
<td>0.029</td>
<td>&lt; 0.001</td>
<td>0.289</td>
<td>0.266</td>
</tr>
<tr>
<td>never_smoker</td>
<td>0.002</td>
<td>0.089</td>
<td>0.017</td>
<td>0.058</td>
<td>&lt; 0.001</td>
<td>0.201</td>
<td>0.377</td>
</tr>
<tr>
<td>age</td>
<td>&lt; 0.001</td>
<td>0.299</td>
<td>0.311</td>
<td>0.045</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>male</td>
<td>0.107</td>
<td>0.096</td>
<td>0.321</td>
<td>0.019</td>
<td>&lt; 0.001</td>
<td>0.018</td>
<td>0.032</td>
</tr>
</tbody>
</table>

### Table A-1. P-values of Coefficients from Multi-Task Learning Methodology

<table>
<thead>
<tr>
<th>Variable</th>
<th>CKD</th>
<th>ARF</th>
<th>IHD</th>
<th>PVD</th>
<th>CHF</th>
<th>CVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1c</td>
<td>0.033</td>
<td>&lt; 0.001</td>
<td>0.132</td>
<td>0.023</td>
<td>0.164</td>
<td>0.007</td>
</tr>
<tr>
<td>fdl</td>
<td>0.052</td>
<td>&lt; 0.001</td>
<td>0.043</td>
<td>0.101</td>
<td>0.289</td>
<td>0.045</td>
</tr>
<tr>
<td>hdl</td>
<td>0.088</td>
<td>0.014</td>
<td>0.101</td>
<td>0.033</td>
<td>&lt; 0.001</td>
<td>0.024</td>
</tr>
<tr>
<td>trigl</td>
<td>0.333</td>
<td>0.010</td>
<td>0.023</td>
<td>0.025</td>
<td>0.003</td>
<td>0.162</td>
</tr>
<tr>
<td>tchol</td>
<td>0.269</td>
<td>&lt; 0.001</td>
<td>0.029</td>
<td>0.045</td>
<td>0.239</td>
<td>0.271</td>
</tr>
<tr>
<td>gfr</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.006</td>
<td>&lt; 0.001</td>
<td>0.021</td>
<td>0.019</td>
</tr>
<tr>
<td>gfr_norm</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.007</td>
<td>&lt; 0.001</td>
<td>0.258</td>
<td>0.033</td>
</tr>
<tr>
<td>bmi</td>
<td>0.085</td>
<td>0.102</td>
<td>0.101</td>
<td>0.138</td>
<td>&lt; 0.001</td>
<td>0.184</td>
</tr>
<tr>
<td>pls</td>
<td>0.348</td>
<td>&lt; 0.001</td>
<td>0.121</td>
<td>0.081</td>
<td>&lt; 0.001</td>
<td>0.088</td>
</tr>
<tr>
<td>sbp</td>
<td>0.055</td>
<td>0.005</td>
<td>0.024</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.007</td>
</tr>
<tr>
<td>dbp</td>
<td>0.154</td>
<td>0.004</td>
<td>0.117</td>
<td>&lt; 0.001</td>
<td>0.138</td>
<td>0.159</td>
</tr>
<tr>
<td>never_smoker</td>
<td>0.344</td>
<td>&lt; 0.001</td>
<td>0.016</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.128</td>
</tr>
<tr>
<td>age</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>0.021</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>male</td>
<td>0.347</td>
<td>&lt; 0.001</td>
<td>0.059</td>
<td>&lt; 0.001</td>
<td>0.002</td>
<td>0.041</td>
</tr>
</tbody>
</table>

### Table A-2. P-values of Coefficients from Baseline Methodology
The Ad-Hoc Uncertainty Principle of Patient Privacy

Jeffrey G. Klann, PhD\textsuperscript{1,2,3}; Matthew Joss, MS\textsuperscript{2}; Rohan Shirali, MA\textsuperscript{5}; Marc Natter, MD\textsuperscript{1,4}; Sebastian Schneeweiss, MD, ScD\textsuperscript{2,3}; Kenneth D. Mandl, MD, MPH\textsuperscript{1,5}; Shawn N. Murphy, MD, PhD\textsuperscript{1,2,3}

\textsuperscript{1}Harvard Medical School; \textsuperscript{2}Partners Healthcare; \textsuperscript{3}Massachusetts General Hospital; \textsuperscript{4}Brigham and Women’s Hospital; \textsuperscript{5}Computational Health Informatics Program, Boston Children’s Hospital. – All in Boston, MA

Abstract

The Health Information Portability and Accountability Act (HIPAA) allows for the exchange of de-identified patient data, but its definition of de-identification is essentially open-ended, thus leaving the onus on dataset providers to ensure patient privacy. The Patient Centered Outcomes Research Network (PCORnet) builds a de-identification approach into queries, but we have noticed various subtle problems with this approach. We censor aggregate counts below a threshold (i.e. <11) to protect patient privacy. However, we have found that thresholded numbers can at times be inferred, and some key numbers are not thresholded at all. Furthermore, PCORnet’s approach of thresholding low counts introduces a selection bias which slants the data towards larger health care sites and their corresponding demographics. We propose a solution: instead of censoring low counts, introduce Gaussian noise to all aggregate counts. We describe this approach and the freely available tools we created for this purpose.

Introduction

Patient Privacy in “De-identified”, Aggregate Data

The Health Information Portability and Accountability Act (HIPAA) allows for the exchange of fully de-identified patient data with many fewer restrictions than data in which patient identity can be determined. Therefore, with the proliferation of large-scale clinical data research networks (e.g. in PCORnet\textsuperscript{[1]} and ACT\textsuperscript{[2]}), transmitting de-identified data is an ideal way to quickly assess study feasibility without being slowed down by regulatory approvals regarding patient privacy.

Unfortunately, assuring that any data set is de-identified is extremely difficult. \textsuperscript{[3]} HIPAA defines 18 distinct identifiers that must be removed from data to ensure it has been de-identified. However, the 18th identifier is “other unique identifying numbers, characteristics or codes,” thus leaving the meaning of de-identification virtually open ended and defined by the ability of a clever adversary to re-identify patient data.

Therefore, even data that is presented in aggregate and not at the patient level could be identifiable if various information in the aggregate data can be combined with personal or public knowledge to re-identify a single patient. Even more insidious is when an aggregate count of patients is counting only one patient. For example, if the number of patients hospitalized greater than six times in the last three months who are black, transgender, and have AIDS equals one, then we know not only enough information to identify the patient but also various demographics about them.

In order to deal with this potential lapse in the protection of patient information, PCORnet has thus far taken a censoring, or thresholding, approach, censoring all aggregate counts in a typical query report that are < 11 to be replaced by a ‘T’. Because individual patients are difficult to identify from sufficiently large aggregate patient counts, one way to prevent patient identification is simply to censor all counts smaller than a predetermined value.

When responding to PCORnet queries, institutions are given the option to censor aggregate counts below a certain threshold. The PCORnet Data Committee has standardized this threshold to <11 patients, but the threshold can be manually adjusted by site researchers prior to submission of results. With the threshold at <11, for example, all counts between 1-10 patients are replaced by a “T”.

Inspection of the results of recent queries has revealed a number of instances where censoring to a threshold is not enough to adequately mask the small aggregate patient counts. In some cases, although small counts are properly censored by a “T”, the “T” can be easily calculated to the exact value below the threshold.

In this manuscript, we analyze situations where this occurs and propose a solution that offers advantages to all parties. This solution is what we dub “The Ad-Hoc Uncertainty Principle of Patient Privacy.”
Methods

Problems with a Censoring Approach

In some cases, although small counts are properly censored by a “T”, they can be easily calculated to the exact value using other available counts. We have seen this in two particular following situations.

Attrition Tables, in which inclusion and exclusion criteria are provided and show the exact number of patients excluded and remaining at each step of the cohort identification process [see Table 1]. In this example, by subtracting the excluded patients from the remaining patients from the previous step, the value of T can be determined. So, for example, the T in Sample1 is 100-98=2 patients.

Patient Characteristics Tables, which display the demographic distribution of the base patient population [see Table 2]. Because all values in each enumerated criteria list are shown, a single T can be inferred by taking the total population count and subtracting every visible criteria count. So, for example, the T in Sample1 is 100-50-49-0=1 patient.

In certain query results, PCORnet query results are not obfuscated at all, as part of an attempt to understand whether variation in distributions may be due to population variability or population size. While PCORnet does not intend to use individual site results (only results aggregated across all sites), these are still sent to the PCORnet Coordinating Center without any low cell count masking, thus potentially exposing patient identity. While PCORnet’s study goals are important, protecting patient information is paramount.

<table>
<thead>
<tr>
<th>Prevalent Event of Interest</th>
<th>Order of Exclusions</th>
<th>Criteria</th>
<th>Remaining Patient Counts</th>
<th>Excluded Patient Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample1</td>
<td>1</td>
<td>Initial Patient Count</td>
<td>100,100</td>
<td></td>
</tr>
<tr>
<td>Sample1</td>
<td>2</td>
<td>Some inclusion/exclusion criteria</td>
<td>100</td>
<td>100,000</td>
</tr>
<tr>
<td>Sample1</td>
<td>3</td>
<td>Some inclusion/exclusion criteria</td>
<td>T</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 1: Attrition Table Example, in which it is possible to infer T.

<table>
<thead>
<tr>
<th>Patient Counts by Characteristics</th>
<th>Sample1</th>
<th>Sample2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall (N)</td>
<td>100</td>
<td>100,000</td>
</tr>
<tr>
<td>By Age Group:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>99</td>
<td>50,000</td>
</tr>
<tr>
<td>25-35</td>
<td>T</td>
<td>49,000</td>
</tr>
<tr>
<td>36-50</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>50+</td>
<td>0</td>
<td>999</td>
</tr>
<tr>
<td>By Sex:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguous</td>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>Male</td>
<td>50</td>
<td>99,999</td>
</tr>
<tr>
<td>Female</td>
<td>49</td>
<td>T</td>
</tr>
<tr>
<td>Other/Missing</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2: Patient Count by Characteristics, in which it is possible to infer the Ts.
Proposed Solution

Various solutions can remedy these problems while retaining the ability to study exact counts of aggregate numbers, including a round-robin data aggregation approach [4] and a homomorphic encryption approach [5]. However, since PCORnet is interested in using these queries to study population trends, rather than specific numbers, we recommend an obfuscation approach utilizing Gaussian noise [6]. This is far easier to implement (both technically and logistically) than other approaches.

Specifically, we propose that PCORnet obfuscate low aggregate counts by adding Gaussian noise to all categorical variable results - a random integer based off of the Normal Gaussian distribution with a mean of 0 and a standard deviation of 2.5 for categorical variables. Refer to Figure 1 below to see the probability distribution for possible random integers this discrete Gaussian could produce. This procedure entirely removes the need to mask low aggregate counts with a “T”.

---

**Figure 1.** Probability distribution for the discrete normalized Gaussian that we propose to use as noise.

Gaussian noise has been previously recommended as an obfuscation technique to mask patient identities [6]. In this paper, Murphy et al. demonstrated that a Gaussian probability distribution was superior to constant or triangular-shaped distribution being that it took longer for the running average of patient counts to converge on the true mean during their Monte Carlo simulations. These simulations represent a simulated attack, where a hacker runs the same query multiple times while taking the running average of these counts until the running average consistently stays within +/- 0.5 of the true patient count. They found that while using a standard deviation of 1.33, it took on average about 12.3 repeats before the running average converged on the true value. Using python, we repeated this simulation using varying standard deviations. We found that using a standard deviation of 2.5, an attacker would need to issue on average about 24 queries before the running average of these counts would converge on the true value. We find this to be a reasonable obstacle to prevent patient identification within PCORNet centered queries, because a query would need to be repeated on the order of 20 times before there is any likelihood of getting the true value. It is unlikely that the PCORNet Query Tool, even with multiple queries potentially interacting with the same subpopulation, will ever reach convergence. (This is especially true considering that data sets are refreshed quarterly, at which point the actual count will change.) Other applications with much higher potential query repeatability might require a different value.
There is a well-understood and standard procedure for adding numbers with associated Gaussian uncertainties. The measurement uncertainties (i.e. uncertainties due to fuzzing) must be propagated down to the aggregate sum.

Error sum propagation can be calculated as follows. For a sum \( f \) of independent counts \( A \) and \( B \) where:

\[
f = A + B
\]

Then the corresponding square of the uncertainty \( \sigma_f^2 \) for this sum is:

\[
\sigma_f^2 = \sigma_A^2 + \sigma_B^2
\]

Therefore, for \( x \) sites that use the same fuzzing uncertainty, the uncertainty of the sum of their counts is equal to the square root of the number of sites multiplied by the Gaussian uncertainty \( \sigma_A \):

\[
\sigma_f = \sigma_A \sqrt{x}
\]

The patient count uncertainty grows with the square root of the number of sites in the final aggregate sum.

Results

We have written a SAS fuzzing script designed to be applied to SAS data sets resulting from queries to the PCORNet data marts. This script takes the following approach:

1) Low counts below a certain threshold (i.e. 11) are identified on all of the data output by a PCORnet analytic query, after the query analysis (so all of the analysis is already completed on the real counts before this occurs).

2) The low counts’ corresponding distributions such as their percentile and cumulative distributions are omitted using a similar method to PCORnet’s current low cell count masking (i.e. replacing the number with ‘T’). These must be omitted since the stratification of patients could narrow results to a single individual, particularly in cases of very low aggregate counts.

3) All of the remaining counts are fuzzed using the Gaussian fuzzing technique. Specifically, for each aggregate count in the data set, the computer picks a randomly-generated number based off of a Gaussian distribution with a mean of ‘0’ and a standard deviation of ‘2.5.’ That randomly generated number is rounded to the nearest integer, and then added to that aggregate count. This is repeated for each aggregate count until every aggregate count has been ‘fuzzed.’

Using this approach, no aggregate counts and only the sensitive distributions corresponding to low cell counts would be omitted. This is in contrast to the current approach, where counts <11 are omitted.

Table 3 shows a version of Table 2 that has been ‘fuzzed’ using this approach rather than thresholded. There are no thresholds because the exact count is no longer presented. Although the categories do not sum to exactly 100% of patients, the technique allows all cells to have a value without risking patient privacy and without implicitly revealing small cell counts, as was the case in Table 2.

This code is available in our GitHub repository, at [https://github.com/ARCH-commons/arch-utils](https://github.com/ARCH-commons/arch-utils). [7]
### Table 2: Patient Count by Characteristics, from Table 2, in which Gaussian noise is applied rather than thresholds.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Sample1</th>
<th>Sample2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall (N)</strong></td>
<td>101</td>
<td>100,005</td>
</tr>
<tr>
<td><strong>By Age Group:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-24</td>
<td>103</td>
<td>50,001</td>
</tr>
<tr>
<td>25-35</td>
<td>2</td>
<td>49,001</td>
</tr>
<tr>
<td>36-50</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>50+</td>
<td>1</td>
<td>996</td>
</tr>
<tr>
<td><strong>By Sex:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ambiguous</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Male</td>
<td>47</td>
<td>100,001</td>
</tr>
<tr>
<td>Female</td>
<td>50</td>
<td>-1</td>
</tr>
<tr>
<td>Other/Missing</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Exact counts of small cells are no longer possible to infer, and all cells have a value.

**Discussion**

As an illustrative analogy, we will compare our ad hoc rule to the Heisenberg uncertainty principle from the field of quantum mechanics. This principle states that it is impossible to simultaneously know the exact location and exact velocity of a particle such as an electron. In this vein of thought, we propose that in all aggregate counts returned to the coordinating center, an observer should not be able to simultaneously determine a count’s exact value as well as that count’s exact query term information. We suggest that PCORnet queries respect this **Ad Hoc Uncertainty Principle of Patient Privacy**. To re-iterate, the Heisenberg Principle insists that the uncertainty in particle position multiplied by the uncertainty in particle momentum is constant, however if one is sure that the particle is in a particular position state (aka an eigenstate of position), then the uncertainty is entirely attributed to the particle’s momentum. Analogously, our proposed rule requires that in order to know the exact query terms for a set of counts (namely the query is in an ‘eigenstate’ of a particular set of query terms), then an uncertainty must be attributed solely to the numeric count, and this uncertainty must be no smaller than 2.5 (the standard deviation of our Gaussian distribution).

This analogy is not perfect. In our example, the uncertainty in the patient counts is finite (2.5) whereas in the case of eigenstates in quantum mechanics the complementary variable that is not certain has an infinite uncertainty. Regardless of this nuance, it is helpful to think of our proposed technique as a way of adding uncertainty to our data so that the identities of patients cannot be discovered, analogously to how the positions of elementary particles are obscured via the Heisenberg Uncertainty Principle.

Analysis of data that has gone through our obfuscation process is akin to the analysis of data with measurement uncertainty in physical science: When a physical scientist makes a measurement, there is an associated measurement uncertainty attributed to that number. This uncertainty is also called the ‘precision’ of the measurement. If the scientist were to then gather a collection of independently measured data points measuring the same observable quantity, this collection would make up a distribution of results which could be Gaussian in shape, with a standard deviation equal to the measurement uncertainty or precision of the individual data points. While our fuzzed counts are not fuzzed due to true measurement uncertainty (and it should not be conflated as such), the aggregation of these counts utilizes a similar process to how a physical scientist would aggregate counts: this is easily accounted for, and networks would be able to accurately report aggregated results from all sites instead of only from a more limited pool of sites with larger counts.

**Advantages over Thresholding**

The PCORnet Coordinating Center has recommended a solution of raising the threshold value so as not to create calculable \( T \) values. However, in order to adequately obfuscate the data, the low cell count threshold must often be increased substantially to the point of obfuscating the results of the query in their entirety. In Table 1, a threshold
value of 99 is needed, and in Table 2, a threshold value of 99,999 is necessary! Our proposed solution would prevent sites from having to increase the obfuscation threshold to the point of obscuring the entire results of a query.

Masking small counts leads to underrepresentation of small databases in studies. When thresholding hides the majority of usable data at a site, then data from small sites with unique demographic representations will be underrepresented, and this will introduce a selection bias slanted toward larger sites and their corresponding demographic pools. The Gaussian-noise obfuscation approach enables smaller sites to contribute their data, no matter how small, which in turn will enable networks to give a more complete and thorough analysis of the data across the network, and thereby providing more far-reaching and useful conclusions for researchers.

Limitations and Future Directions
In some circumstances where all sites have low counts, the current thresholding technique is superior to Gaussian noise. In particular, when the aggregate sum is less than or about equal to the aggregate uncertainty for a particular number of sites, then a thresholded censorship approach is best. This is illustrated in Figure 2 below. We plot the aggregate uncertainty (shaded area) versus the aggregate sum. The sums are represented by a set of straight lines with a slope equal to the average number of counts per site, where ‘x’ is the number of sites. When the average site reports a count of 2.5 or more for a query (i.e. slope is greater than or equal to 2.5), then the aggregate uncertainty is never greater than the aggregate sum, and therefore these cases are all well suited for our Gaussian fuzzing technique. In contrast, for smaller slopes such as 1 or 0.5, it may be better for the network to use thresholded censoring for smaller numbers of sites. When the aggregate sum is less than or about equal to the aggregate uncertainty for a particular number of sites, such as when x is less than or equal to 6 for the aggregate sum function with a slope of 1, then a thresholded censorship approach is best.

![Figure 2: Aggregate sums represented by linear functions where the slope signifies the average number of counts per site and where x represents the number of sites. These aggregate sum functions are compared to the aggregate uncertainty (shaded area).](image)

Using this method of Gaussian fuzzing to obfuscate our results has an interesting quirk: one needs to potentially accept negative values for counts as they may be necessary in order to preserve the correct fuzzing uncertainty from each individual site. Thus, by extension these negative values are necessary in order to assess the correct fuzzing uncertainty after error propagation. To understand this, consider a query looking for counts of patients with a rare disease. Suppose
in one case, a patient count is ‘3’ before fuzzing. Our proposed method of introducing Gaussian noise would then produce a random number based off of the Gaussian distribution with a mean of ‘0’ and a standard deviation of ‘2.5,’ which could produce a negative number to be summed with the original count. This could in turn cause the fuzzed count to become a negative number, such as ‘-2.’ This negative number must be retained in order to preserve the Gaussian properties of the cumulative distribution; if it were rejected, the distributions of counts generated after fuzzing would be skewed to be larger than expected and may no longer be Gaussian, and thus the aforementioned error propagation formulas described in the Results section will no longer be valid.

Any queries that result in ‘0’ counts do not need to be fuzzed to protect patient privacy, which could be accounted for in order to reduce the aggregate uncertainties. However, this would also require an additional reference table to keep track of fuzzed versus non-fuzzed zeros.

In the future, it may be useful to examine how this technique could work in concert with a query lockout feature that would inhibit a user from issuing the same query too many times within a certain period. This could be implemented not only to prevent Gaussian convergence of the true number of patients, but also if the lockout threshold is low enough (e.g. 10), then the fuzzing uncertainty could also be decreased to an appropriate value (e.g. 1.33, which would take about 12.3 tries to converge on the mean). This could enable smaller sites or sites with small counts to undergo less fuzzing, thereby enabling their aggregate counts to surpass their aggregate uncertainties.

**Conclusion**

This approach of ceasing the censorship of low counts while introducing Gaussian noise to all aggregate counts will help data research networks obtain a more complete report that more accurately reflects the data across the entire network, avoiding potential selection biases slanted towards larger healthcare sites and more prevalent demographics pools.

**Funding**

This work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-04608) for development of the National Patient-Centered Clinical Research Network, known as PCORnet.

**Disclaimer**

The statements presented in this publication are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee or other participants in PCORnet.

**References**


6. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. Proc *AMIA Symp* 2002;552.

Using Machine Learning Algorithms to Predict Risk for Development of Calciphylaxis in Patients with Chronic Kidney Disease

Ross S. Kleiman, M.S.,1,2 Eric R. LaRose, B.S.,1 Jonathan C. Badger, Pharm.D.,1,3 David Page, Ph.D.2,3 Michael D. Caldwell, M.D., Ph.D.4 James A. Clay, M.D.5 Peggy L. Peissig, Ph.D.,1
1Marshfield Clinic Research Institute, Marshfield, WI; 2Department of Computer Sciences, University of Wisconsin, Madison, WI; 3Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI; 4Center for Hyperbaric Medicine and Tissue Repair, Marshfield Clinic; 5Department of Surgery, Marshfield Clinic

Abstract
Calciphylaxis is a disorder that results in necrotic cutaneous lesions with a high rate of mortality. Due to its rarity and complexity, the risk factors for and the disease mechanism of calciphylaxis are not fully understood. This work focuses on the use of machine learning to both predict disease risk and model the contributing factors learned from an electronic health record data set. We present the results of four modeling approaches on several subpopulations of patients with chronic kidney disease (CKD). We find that modeling calciphylaxis risk with random forests learned from binary feature data produces strong models, and in the case of predicting calciphylaxis development among stage 4 CKD patients, we achieve an AUC-ROC of 0.8718. This ability to successfully predict calciphylaxis may provide an excellent opportunity for clinical translation of the predictive models presented in this paper.

Introduction
Calciphylaxis, also known as calcific uremic arteriolopathy, is a highly morbid disorder that presents with necrotic lesions of the skin resultant at least partially from ischemia caused by calcification of the small and medium-sized arteries. The mortality rate of calciphylaxis has been measured in excess of 50% within one year of initial diagnosis. One of the most commonly associated comorbidities with calciphylaxis is chronic kidney disease (CKD). In particular, calciphylaxis is found most often among patients in the late stages of CKD and end-stage renal disease (ESRD). While calciphylaxis was originally named by Hans Selye in 1962, there is still much that is unknown about its origin and risk factors. Known risk factors include: CKD, ESRD, mineral and bone disorders, diabetes mellitus, hyperphosphatemia, female gender, obesity, warfarin use, and ethnicity.

Currently, calciphylaxis is identified by histological examination of excised tissue. This examination and confirmation of calciphylaxis occurs after a patient has already developed the clinical manifestations of this highly morbid disease. Currently, there are no risk assessment models used in clinical practice for identifying patients at risk for development of calciphylaxis. In this work, we aim to produce such a risk assessment model for patients with CKD. A diagnosis of CKD is accompanied by regular monitoring through frequent patient contacts with their healthcare system. The translation of a calciphylaxis risk assessment model targeted to patients with CKD could allow for intervention which may prevent the disease.

In recent years, the use of electronic health records (EHRs) has increased in healthcare systems. This wealth of data has created unprecedented opportunities for computational approaches to not only augment existing knowledge of various diseases but also produce predictive models to assess patient risk. Diseases such as breast cancer and myocardial infarction have already been successfully modeled using machine learning algorithms. Machine learning, a branch of artificial intelligence, focuses on producing algorithms that learn rules or relationships about a set of variables to predict the value or outcome of an unknown variable. In the context of healthcare data, machine-learned models can be used to predict the disease risk of a patient from the information present in their health record. In some cases, the models produced by a machine learning algorithm can be manually inspected to understand which variables suggest different patient outcomes. In this paper, we present the use of two machine learning algorithms, lasso-penalized logistic regression and random forests, in the context of prediction and risk factor analysis for calciphylaxis. We present these methodologies on several subpopulations of CKD and ESRD and show that they produce a strong prediction of calciphylaxis.

Experimental Methodology

Data
Our dataset is derived from patient visits to the Marshfield Clinic, a health care system that serves Northern and Central Wisconsin. Data include diagnosis codes, in the forms of both ICD-9 and ICD-10 codes, demographic information, laboratory values, vital sign readings, procedures, and medications present on patient records. Apart from age and gender, all features were initially extracted as counts, that is the number of times an event occurred on a patient record from the start of their record to the censor date which we later define. These values were gathered from patients identified as case or control patients for our experiments. Case patients were manually identified by their attending surgeon, who confirmed calciphylaxis diagnosis via inspection of patient records and histological examination of excised tissue. Control patients were required to have no diagnosis codes indicative of calciphylaxis in their records including ICD-9 codes 275.49 and 709.1, and ICD-10 codes L95.9 and E83.59. Moreover, for both case and control patients, we required a minimum of 10 unique diagnosis codes on their records. The minimum data requirement was implemented to ensure sufficient patient history for prediction.

**Case-Control Matching**

We present research on 5 different CKD patient populations: patients with any stage of CKD, which we refer to as “any stage”, patients with stage 3 CKD (stage 3), patients with stage 4 CKD (stage 4), patients with stage 5 CKD (stage 5), and patients with stage 5 CKD requiring chronic dialysis, which we refer to as end-stage renal disease (ESRD). In all experiments, we attempted to match up to 10 control patients for each case patient. We chose a 10-to-1 ratio based on the small number of confirmed cases. For the purposes of illustrating our case-control matching scheme consider two patients, Bob (case) and Alex (control), who will become an eventual case-control pair. Bob was manually identified as a case patient via histological examination and inspection of his patient record. Furthermore, we require Bob to have some stage of CKD and place him in the appropriate population. We then search for candidate control patients to match with Bob who have a birthdate within 30 days of Bob’s and the same gender. Let Alex be one of these candidate control patients. Alex must have the same stage of CKD as Bob on or before Bob’s calciphylaxis diagnosis date. Furthermore, Alex must have entries on his record both before and after Bob’s calciphylaxis diagnosis date. This control data straddling of the case diagnosis date allows us to ensure that the control patient was alive and present in the Marshfield system during the case patient’s diagnosis. Now that we have identified Alex as a control match for our case patient, Bob, we finally truncate both their patient records by removing all data following 30-days prior to Bob’s calciphylaxis diagnosis date. This truncation allows us to control for class-label leakage and ensure we are performing a prediction task. This case-control matching procedure was performed up to 10 times for each case patient. Details of these patient populations are in Table 1.

**Table 1.** Description of the five experimental patient populations. For each population, we describe the ICD-9 and ICD-10 diagnosis codes associated. Moreover, we present the number of case and control patients for each population and the total number of non-zero features on their records that were included in the model.

<table>
<thead>
<tr>
<th></th>
<th>Any Stage</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
<th>ESRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD-9 Code(s)</td>
<td>585-585.9</td>
<td>585.3</td>
<td>585.4</td>
<td>585.5</td>
<td>585.6</td>
</tr>
<tr>
<td>ICD-10 Code(s)</td>
<td>N18-N18.9</td>
<td>N18.3</td>
<td>N18.4</td>
<td>N18.5</td>
<td>N18.6</td>
</tr>
<tr>
<td># Features</td>
<td>9,288</td>
<td>5,037</td>
<td>6,662</td>
<td>5,864</td>
<td>6,974</td>
</tr>
<tr>
<td># Cases</td>
<td>38</td>
<td>10</td>
<td>15</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td># Controls</td>
<td>363</td>
<td>100</td>
<td>148</td>
<td>117</td>
<td>165</td>
</tr>
<tr>
<td><strong>Total Patients</strong></td>
<td>401</td>
<td>110</td>
<td>163</td>
<td>129</td>
<td>182</td>
</tr>
</tbody>
</table>

**Feature Construction**

In addition to our experimentation based on the stage of chronic kidney disease, we explored the effects of binary versus continuous feature representations. The raw dataset was structured such that each non-demographic variable, e.g., a lab test, had a value for the number of times the patient had received that particular variable on their record. The counts for each variable were done from the start of the patient’s record until an experiment-specific date, e.g., their first entry of a particular stage of chronic kidney disease. One potential concern about using count data is the likelihood of chronically sick patients to have an increased number of healthcare encounters. In this way, uncorrelated variables that are recorded often such as a height measurement vital could become inappropriately correlated with
disease status. This was of particular concern as calciphylaxis tends to occur in much sicker patients. For this reason, we also explored the use of binary features. We constructed these features by setting any non-demographic feature value to 1 if the patient had ever had it entered on their record, and 0 if they had never had it on their record. In the case of continuous features, we simply scaled the counts for each feature to be between 0 and 1, which was of use when comparing the relative importance of features when evaluating our models.

**Model Construction Methods**

We explored the predictive capacities of two different machine learning models: random forests and logistic regression. Random forest models are well known for both their strong accuracy and their resilience to high dimensional data. Furthermore, random forests can capture nonlinear interactions in data. While random forests are strong predictors, they are often difficult to interpret, and interpretation is often critical in a healthcare context. For this reason, we additionally explored logistic regression models. Because logistic regression is not inherently resilient to high dimensional data, we employed lasso-penalized logistic regression, which utilizes an L1-regularization term in the objective function of the model to penalize features that were only marginally informative and to encourage a small set of strongly predictive features in the final model. This is necessary not only to obtain comprehensible models but also to obtain accurate ones, as the patient populations we modeled included thousands of unique features and logistic regression without some form of dimensionality reduction can perform very poorly in such situations. The L1 or “lasso” penalty is a widely-used approach to dimensionality reduction in regression.

For both the random forest and the logistic regression models we performed a case-control matched leave-one-case-out cross-validation. For each experiment, we chose k-folds, one fold for each of the k case patients and its matching controls. In this procedure, for each round of cross validation one fold was chosen to be left out and was comprised of a single case patient and the up to 10 control patients matched with it. In the case of random forests, models were constructed using the remaining k − 1 folds of training data. We built 500 decision trees for each forest and used the square root of the number of features as the number of features to consider at each split. Trees were grown to leaf purity where possible and splits were chosen via Gini gain that was calculated in a balanced way, purity where possible and splits were chosen via Gini gain that was calculated in a balanced way.

For logistic regression, we performed an additional layer of internal leave-one-case-out cross-validation to tune the penalty coefficient for the L1 regularization term. That is, we performed an internal k − 1 rounds of internal cross validation with 1 tuning fold and k − 2 training folds. In this tuning procedure, we consider 10 different penalty values logarithmically spaced between $10^{-4}$ and $10^9$. Each penalty value was evaluated via an internal cross-validation layer and the optimal penalty value was chosen based on the performance of the various models as judged by a weighted accuracy measurement that ensured equal combined weights for the cases and the controls. Thus, for each external cross-validation fold, the remaining folds were used to select a penalty value and a new predictive model was trained on those folds and finally evaluated on the original held out fold. Logistic regression models were all constructed using a balanced class weight approach like that employed in the random forest models.

For each of the 20 choices of CKD population, model family, and feature type, e.g. predicting CKD stage 4 with random forests using binary features, we repeated the model construction 30 times. To account of the skew of our data, for both, random forest and logistic regression we constructed the models with the “balanced” class-weight option in the scikit-learn library. These replications were done because random forests are stochastic by nature, and logistic regression may have multiple equally good solutions to their optimization problem, both of which can produce varying results across multiple runs from the same data set. In this way, we could better estimate the predictive quality achieved under each experimental condition.

**Model Evaluation Methods**

Model evaluation is done quantitatively via the construction of receiver operating characteristic (ROC) curves and precision-recall (PR) curves. While we report both the area under the ROC-curve (AUC-ROC) and PR-curve (AUC-PR), we primarily use the AUC-ROC as a numerical evaluation of the quality of our models. We do this because AUC-ROC provides a metric of efficacy that is unbiased by the class skew of a dataset. Because we attempt to achieve a 10-to-1 control to case ratio this may not be representative of the true population skew, and thus we provide a fairer analysis of our results via ROC-curves. For our best performing model and feature combination, we present both ROC- and PR-curves for completeness but do not use them in ranking the quality of the models. Both ROC-curves and PR-curves were constructed for each model in a similar fashion. For a single repetition of a single experiment we first produced a vector of predicted probabilities corresponding to model estimated risk for each patient. We produced this vector during the leave-one-case-out k-fold cross validation by using the model built on the training folds to
predict the labels of the test fold. Our final predicted probability vector was the union of these predictions as each patient was only in a test fold once. From these predicted probabilities and the true class labels we constructed either a ROC- or PR-curve in the typical fashion. Then across the 30 repetitions for each experiment, we performed vertical averaging of the 30 resulting curves to yield a final curve.

For each of the five experimental conditions, we wish to know which of the four models performed best. We evaluated model performance via a sign test and evaluated 12 hypotheses which were of the form: “Binary feature random forest is significantly better than binary feature logistic regression”. These twelve hypotheses were all of the unique ordered pairs of the four models drawn without replacement. For each experimental condition, we chose an experiment-wide p-value of $\alpha = 0.05$ as our threshold for significance. We performed a Bonferroni correction and required $p_i \leq \frac{0.05}{12}$ for an individual hypothesis to be significant. If a model was significantly better than the three other models, then we considered it to be the best model for that experimental condition.

Our p-values for each hypothesis were calculated via the aforementioned sign test, the details of which we provide now. Because the 30 repetitions of each experimental condition and model are not truly independent of one another, we cannot use bootstrapping to calculate confidence intervals on our AUCs. Instead, we use a sign test to determine if a model is correct more often than another model a statistically significant number of times. For each model, we first construct a ROC-curve in the fashion described above. To convert the probabilities from each algorithm to labels, we choose a threshold corresponding to an 80% true positive rate, or recall, for our algorithms. Because of the serious nature and high mortality rate of calciphylaxis we chose 80% recall to reflect the significant cost difference between false negatives and false positives. Thus, for each experiment, the threshold corresponding to 80% recall was used to label predictions as either positive or negative. For every patient we then checked the label assigned by both models and its true label, if both models assigned the same label, that patient was considered a tie and not counted, however, if the models differed then the first model received either a win or a loss if it was correct or incorrect respectively. Then, each hypothesis was tested by first counting $n_{\text{wins}}$, the number of times the first model predicted a patient correctly and the second model predicted that patient incorrectly, and $n_{\text{loss}}$, the converse statement. Finally, we calculate the cumulative distribution of the binomial function $\text{Binom}(n_{\text{wins}}, n_{\text{loss}}, 0.5)$ to determine a p-value for a particular hypothesis.

We perform a qualitative evaluation of our models by inspecting the top performing features returned from our models. In the case of random forests, we used the approach detailed in Breiman 1984 to determine feature importance values via the Gini importance.¹⁰

**Results**

In general, we find that we achieve superior performance predicting calciphylaxis from a specific stage of CKD compared to an arbitrary stage. We present in Table 2 the precise ROC-AUC values achieved in each experiment. We note that in general random forests outperform logistic regression and that use of binary features roughly outperforms the use of continuous features. In two of the five experimental conditions, we found random forests with binary features to outperform the other three approaches with a statistically significant difference.

Given the strong performance of random forests using binary features, we chose to visualize the ROC- and PR-curves for this model and feature engineering scheme across the five experimental conditions and present this in Figure 1. We note that the models predicting CKD at a specific stage appear to be clustered together and roughly separated from the ROC-curve representing prediction at an arbitrary stage of CKD. Additionally, there appears to be some separation between the PR-curve representing prediction of ESRD as compared to the other stages of CKD.
Table 2. AUC-ROC, AUC-PR, Accuracy, and Precision @ Recall = 80% values for each experimental condition and algorithm. For each experimental condition, the algorithm that performed best is highlighted in bold if it is statistically significantly better than the other three algorithms as determined by a sign test. Note that just binary random forests achieved statistical significance and did so in two experimental conditions. For CKD stage 4 binary random forests outperformed its competitors with $p \leq 3.78e - 04$. For ESRD stage 4 binary random forests outperformed its competitors with $p \leq 1.92e - 04$.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Model</th>
<th>Features</th>
<th>AUC-ROC</th>
<th>AUC-PR</th>
<th>Accuracy</th>
<th>Pre. @ 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Logistic Regression</td>
<td>Binary</td>
<td>0.789</td>
<td>0.189</td>
<td>0.545</td>
<td>0.190</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.785</td>
<td>0.215</td>
<td>0.545</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>Binary</td>
<td>0.846</td>
<td>0.261</td>
<td>0.545</td>
<td>0.212</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.847</td>
<td>0.279</td>
<td>0.545</td>
<td>0.230</td>
</tr>
<tr>
<td>4</td>
<td>Logistic Regression</td>
<td>Binary</td>
<td>0.791</td>
<td>0.190</td>
<td>0.515</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.635</td>
<td>0.125</td>
<td>0.528</td>
<td>0.131</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>Binary</td>
<td>0.872</td>
<td>0.292</td>
<td>0.546</td>
<td>0.269</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.791</td>
<td>0.200</td>
<td>0.546</td>
<td>0.194</td>
</tr>
<tr>
<td>5</td>
<td>Logistic Regression</td>
<td>Binary</td>
<td>0.707</td>
<td>0.194</td>
<td>0.543</td>
<td>0.164</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.689</td>
<td>0.169</td>
<td>0.532</td>
<td>0.195</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>Binary</td>
<td>0.796</td>
<td>0.212</td>
<td>0.535</td>
<td>0.248</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.770</td>
<td>0.184</td>
<td>0.535</td>
<td>0.213</td>
</tr>
<tr>
<td>ESRD</td>
<td>Logistic Regression</td>
<td>Binary</td>
<td>0.746</td>
<td>0.168</td>
<td>0.538</td>
<td>0.168</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.750</td>
<td>0.198</td>
<td>0.549</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>Binary</td>
<td>0.858</td>
<td>0.440</td>
<td>0.538</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.838</td>
<td>0.425</td>
<td>0.538</td>
<td>0.244</td>
</tr>
<tr>
<td>Any</td>
<td>Logistic Regression</td>
<td>Binary</td>
<td>0.713</td>
<td>0.188</td>
<td>0.534</td>
<td>0.156</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.734</td>
<td>0.162</td>
<td>0.541</td>
<td>0.173</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>Binary</td>
<td>0.738</td>
<td>0.199</td>
<td>0.536</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Continuous</td>
<td>0.725</td>
<td>0.185</td>
<td>0.536</td>
<td>0.156</td>
</tr>
</tbody>
</table>
Figure 1. ROC- and PR-Curves for binary feature random forest across the five experimental conditions.

Table 3. Ranking of 10 top random forest model features for each of the 5 CKD experiments: any CKD stage, stage 3 CKD, stage 4 CKD, stage 5 CKD, and ESRD. For each experiment, feature importance values were averaged first across each of the $k$-folds for cross-validation, and then those values were averaged across the 30 repetitions completed to produce a final importance value for each feature.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Any Stage</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
<th>ESRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hepatitis B Surface Ag</td>
<td>Obesity</td>
<td>Anemia in CKD</td>
<td>Obesity</td>
<td>Thyroxine (T4)</td>
</tr>
<tr>
<td>2</td>
<td>Anemia in CKD</td>
<td>Lactescence/Chylomicrons</td>
<td>Morbid Obesity</td>
<td>Morbid Obesity</td>
<td>Obesity</td>
</tr>
<tr>
<td>3</td>
<td>Secondary Hyper-parathyroidism of Renal Origin</td>
<td>Chylomicrons</td>
<td>Thyroid Stimulating Hormone-Reg'l C</td>
<td>Parathyroid Hormone (PTH), 1-84</td>
<td>Amylase-Pancreatic</td>
</tr>
<tr>
<td>4</td>
<td>Direct Microscopy</td>
<td>ALT (GPT)</td>
<td>Obesity</td>
<td>Radiologic exam knee complete 4/more views</td>
<td>Age</td>
</tr>
<tr>
<td>5</td>
<td>Ulcer of Lower Limb, Unspecified</td>
<td>Differential Polychromatophili</td>
<td>Chylomicrons</td>
<td>Instrument Neutrophil #</td>
<td>Frac.O2 Hb, Arterial</td>
</tr>
<tr>
<td>6</td>
<td>Iron Defic Anemia Nos</td>
<td>Differential Poikilocytosis</td>
<td>Lactescence/Chylomicrons</td>
<td>Age</td>
<td>Arthropathy, unspecified</td>
</tr>
<tr>
<td>7</td>
<td>Hepatitis B Surface (HBs) Ab</td>
<td>Uric Acid, Blood</td>
<td>Thyroxine (T4)</td>
<td>Uric Acid, Bld</td>
<td>Prothrombin Time (PT)</td>
</tr>
<tr>
<td>8</td>
<td>% O2 Saturation</td>
<td>Differential Activated Lymph</td>
<td>Secondary Hyper-parathyroidism of Renal Origin</td>
<td>Non-HDL Cholesterol</td>
<td>Abdominal Pain</td>
</tr>
<tr>
<td>9</td>
<td>Differential Poikilocytosis</td>
<td>Sodium serum plasma or whole blood</td>
<td>Prescription transmit via erx system</td>
<td>Prescription transmit via erx system</td>
<td>Chronic Liver Disease Nec</td>
</tr>
<tr>
<td>10</td>
<td>Blood Urea Nirtrogen-Post-Dial</td>
<td>Platelet Estimate</td>
<td>Hypercholesterolemia</td>
<td>Skin Suture Nec</td>
<td>Blood count complete automated</td>
</tr>
</tbody>
</table>
For each of the five experimental conditions, we present in Table 3 the top 10 features used in prediction for the binary feature random forest. We note that many of these features are strongly related to many already known risk factors for calciphylaxis. Moreover, these features are suggestive that we successfully control for CKD, as few CKD related features appear among the top features for these models. We find that there may be some temporal confounding, as the feature “Prescription transmit via erx system” is present in the top ten features for both stage 4 and stage 5 models. This particular feature can appear when a model is attempting to distinguish cases and controls temporally, as the introduction of electronic transmission of prescriptions can be used to infer if a patient’s data is from a more recent period by inspecting if this feature is present on their record.

Discussions
In this work, we explored prediction of calciphylaxis at various stages of CKD, through the use of random forest or logistic regression models and either binary or continuous features. We find that overall prediction of calciphylaxis is achievable with strong results and note that in each of the experimental conditions in which we predict calciphylaxis during a specific stage of CKD, we find at least one model with an AUC-ROC of nearly 0.8 or above. Such AUC-ROC values suggest models of a high quality capable of predicting calciphylaxis with reasonable efficacy. Given the high mortality rate of calciphylaxis, we feel that early prediction of calciphylaxis would be of great benefit for physicians who wish to monitor patient risk. In particular, we find some of our strongest results when predicting calciphylaxis using binary feature random forests for the experimental conditions of stages 3 and 4 where patients have moderate to advanced CKD.

It is worth noting that for a given task and feature type, random forests nearly always outperformed logistic regression. We feel that this is likely due to the ability for random forests to capture nonlinear interactions in data. Furthermore, for a given task and model type, binary features achieved AUC-ROC values that were typically equal to or better than those achieved with the use of continuous features. This is of interest as our constructions for binary and continuous features ensured that continuous features had strictly greater information. For every feature other than age, the binary feature could be obtained from the continuous feature by thresholding such that a continuous feature of value 0 becomes a binary feature of value 0, and a continuous feature of value anything greater than 0 becomes a binary feature of value 1. In the case of random forests, this could be expressed as splitting on a particular continuous feature, \( f_i \), with \( f_i > 0 \) being the logical test for indicating which branch to choose. The poorer performance from continuous features suggests that there may be some overfitting occurring. By using binary features, we are limiting the expressiveness of our models, and such an action could cause improvement in the case of a model that is overfitting.

In our analysis of the top performing features for the various models and experimental conditions, we found that features could be grouped roughly into one of three categories: previously known calciphylaxis risk factors, indicators of potential confounding in the model, or potential calciphylaxis risk factors with minimal or no previous support. Of those risk factors for calciphylaxis that are previously discovered, our models pick up most strongly on obesity and its related risks, anemia/low iron, and hyperparathyroidism. It is worth noting that we see minimal use of CKD features, which suggests that our method for controlling for CKD was largely successful. There are some features which suggest either temporal confounding. In particular, the feature “Prescription transmit via erx system” is likely being used to separate patients by how recently they visited the Marshfield Clinics. Such a feature can be exploited during cases of temporal confounding as the use of an electronic prescription transmittal system is a relatively recent addition to workflows in healthcare systems. Finally, we find some of our top features are related to liver disease and/or liver function, a relationship that is not strongly established with calciphylaxis. In particular, we see both various liver function labs and testing for hepatitis B as strong features for predicting calciphylaxis.

Conclusion
Patients suffering from late-stage CKD and ESRD are known to be at a much higher risk for calciphylaxis, a vascular disorder with a one-year mortality rate in excess of 50%. By identifying at-risk patients, we provide an opportunity for these patients to take actions to mitigate their risk factors and avoid a potentially deadly disease. Additionally, by modeling calciphylaxis to predict patient risk, we invite the opportunity to discover more about the potential risk factors of what is a poorly understood disease. In this work, we explored prediction of calciphylaxis at various stages of CKD, through the use of random forest or logistic regression models and either binary or continuous features. From this work, we found that random forests using binary features tended to outperform other methodologies and that we could typically predict calciphylaxis with strong efficacy. Furthermore, we found that our models not only largely exploited known risk factors for calciphylaxis but also found some evidence for a previously unsupported connection to liver function and hepatitis. We note that our results show these models to be of high quality and intend to explore
the potential for translating these models to a clinical setting, where they could be used to identify patients who are of higher risk for developing calciphylaxis.

Acknowledgements

The authors gratefully acknowledge the developers and maintainers of the programming languages and packages used in the experimentation and evaluation of this work: Python, scikit-learn, numpy, pandas, scipy, matplotlib, and seaborn.

Financial Support

The authors gratefully acknowledge the support of the NIH BD2K grant: U54 AI117924, the NIG NIGMS grant: R01 GM097618, the NLM training grant: 5T15LM007359, and the NIH NCATS grant: UL1 TR000427.

References

Automated Detection of Diabetic Retinopathy using Deep Learning

Carson Lam, MD¹, Darvin Yi¹, Margaret Guo², Tony Lindsey, PhD¹,³
¹Biomedical Informatics Department, Stanford University, Palo Alto, CA; ²School of Medicine, Stanford University, Palo Alto, CA; ³NASA Ames Research Center, Mountain View, CA

Abstract

Diabetic retinopathy is a leading cause of blindness among working-age adults. Early detection of this condition is critical for good prognosis. In this paper, we demonstrate the use of convolutional neural networks (CNNs) on color fundus images for the recognition task of diabetic retinopathy staging. Our network models achieved test metric performance comparable to baseline literature results, with validation sensitivity of 95%. We additionally explored multinomial classification models, and demonstrate that errors primarily occur in the misclassification of mild disease as normal due to the CNNs inability to detect subtle disease features. We discovered that preprocessing with contrast limited adaptive histogram equalization and ensuring dataset fidelity by expert verification of class labels improves recognition of subtle features. Transfer learning on pretrained GoogLeNet and AlexNet models from ImageNet improved peak test set accuracies to 74.5%, 68.8%, and 57.2% on 2-ary, 3-ary, and 4-ary classification models, respectively.

1. Introduction

Approximately four hundred and twenty million people worldwide have been diagnosed with diabetes mellitus. The prevalence of this disease has doubled in the past 30 years²⁴ and is only expected to increase, particularly in Asia⁷. Of those with diabetes, approximately one-third are expected to be diagnosed with diabetic retinopathy (DR), a chronic eye disease that can progress to irreversible vision loss⁸. Early detection, which is critical for good prognosis, relies on skilled readers and is both labor and time-intensive. This poses a challenge in areas that traditionally lack access to skilled clinical facilities. Moreover, the manual nature of DR screening methods promotes widespread inconsistency among readers. Finally, given an increase in prevalence of both diabetes and associated retinal complications throughout the world, manual methods of diagnosis may be unable to keep pace with demand for screening services¹².

Automated techniques for diabetic retinopathy diagnoses are essential to solving these problems. While deep learning for binary classification in general has achieved high validation accuracies, multi-stage classification results are less impressive, particularly for early-stage disease.

In this paper we introduce an automatic DR grading system capable of classifying images based on disease pathologies from four severity levels. A convolutional neural network (CNN) convolves an input image with a defined weight matrix to extract specific image features without losing spatial arrangement information. We initially evaluate different architectures to determine the best performing CNN for the binary classification task and aim to achieve literature reported performance levels. We then seek to train multi-class models that enhance sensitivities for the mild or early stage classes, including various methods of data preprocessing and data augmentation to both improve test accuracy as well as increase our effective dataset sample size. We address concerns of data fidelity and quality by collating a set of ophthalmologist verified images. Finally, we address the issue of insufficient sample size using a deep layered CNN with transfer learning on discriminant color space for the recognition task. We then trained and tested two CNN architectures, AlexNet and GoogLeNet, as 2-ary, 3-ary and 4-ary classification models. They are tuned to perform optimally on a training dataset using several techniques including batch normalization, L2 regularization, dropout, learning rate policies and gradient descent update rules. Experimental studies were conducted using two primary data sources, the publicly available Kaggle dataset of 35,000 retinal images with 5-class labels (normal, mild, moderate, severe, end stage) and a physician-verified Messidor-1 dataset of 1,200 color fundus images with 4-class labels. Throughout this study we aim to elucidate a more effective means of classifying early stage diabetic retinopathy for potential clinical benefits.
2. Background and Related Work

Diagnosis of pathological findings in fundoscopy, a medical technique to visualize the retina, depends on a complex range of features and localizations within the image. The diagnosis is particularly difficult for patients with early stage diabetic retinopathy as this relies on discerning the presence of microaneurysms, small saccular outpouching of capillaries, retinal hemorrhages, ruptured blood vessels—among other features—on the fundoscopic images. Prototypical retinal disease stages are shown in Fig. 1.

Computer-aided diagnosis of diabetic retinopathy has been explored in the past to reduce the burden on ophthalmologists and mitigate diagnostic inconsistencies between manual readers. Automated methods to detect microaneurysms and reliably grade fundoscopic images of diabetic retinopathy patients have been active areas of research in computer vision. The first artificial neural networks explored the ability to classify patches of normal retina without blood vessels, normal retinas with blood vessels, pathologic retinas with exudates, and pathologic retinas with microaneurysms. The accuracy of being able to detect microaneurysms compared to normal patches of retina was reported at 74%.

Past studies using various high bias, low variance digital image processing techniques have performed well at identifying one specific feature used in the detection of subtle disease such as the use of top-hat algorithm for microaneurysm detection. However, a variety of other features besides microaneurysms are efficacious for disease detection.

Additional methods of detecting microaneurysms and grading DR involving k-NN, support vector machines, and ensemble-based methods have yielded sensitivities and specificities within the 90% range using various feature extraction techniques and preprocessing algorithms.

Previous CNN studies for DR fundus images achieved sensitivities and specificities in the range of 90% for binary classification categories of normal or mild vs moderate or severe on much larger private datasets of 80,000 to 120,000 images. However, accuracy measures for the detection of four classes of DR, that is: no DR (R0), mild (R1), moderate (R2), and severe (R3) depend nontrivially on disease graded class collection ratios. While R0 and R3 stages are capable of achieving high sensitivity, the R1 and R2 computed recall rates are often low. Experiments from publicly available datasets suggest this is primarily attributable to the relative difficulty of detecting early stage DR. Furthermore, current accuracies for R1 and R2 stages are reported at 0% and 41%, respectively. We will determine the sensitivity and specificity of our 4-ary classification model and evaluate performance by comparing results to currently published research data.

3. Dataset

We used two fundoscope image datasets to train an automated classifier for this study. Rapid prototyping was facilitated by pretrained models obtained from the ImageNet visual object recognition database. Diabetic retinopathy images were acquired from a Kaggle dataset of 35,000 images with 5-class labels (normal, mild, moderate, severe, end stage) and Messidor-1 dataset of 1,200 color fundus images with 4-class labels (normal, mild, moderate, severe). Both datasets consist of color photographs that vary in height and width between the low hundreds to low thousands. Compared to Messidor-1, the Kaggle dataset consists of a larger proportion of uninterpretable images due to artifact preponderance, faulty labeling and poor quality. After training on the larger Kaggle datasets and identifying limitations of the conventional approach to retinal image classification, we performed experiments on higher fidelity datasets with improved image quality and reliable labeling.
In the interest of efficient model building, we progressed to a smaller but more ideal dataset for learning difficult features. The Messidor dataset was supplemented with a Kaggle partition (MildDR) consisting of 550 images that was verified for its efficacy by direct physician interpretation. The dataset contains images from a disparate patient population with extremely varied levels of fundus photography lighting and is labeled in a consistent manner. The lighting affects pixel intensity values within the images and creates variation unrelated to classification pathology. Our study only uses the retinopathy grade as a reference, a description of which is provided in Table 1, together with the number of images for each category.

To assess the suitability of transfer learning, we considered several frameworks to design, train and deploy a modified version of GoogLeNet as a baseline 2-ary, 3-ary and 4-ary classification model. Our final model was implemented in TensorFlow and was influenced by results from a deep learning GPU interactive training system (DIGITS) that enabled rapid neural network prototype training, performance monitoring and real-time visualizations.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Description</th>
<th>Nb Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0</td>
<td>((N_{MA} = 0) \text{ AND } (N_{HE} = 0))</td>
<td>546</td>
</tr>
<tr>
<td>R1</td>
<td>((0 &lt; N_{MA} \leq 5) \text{ AND } (N_{HE} = 0))</td>
<td>153</td>
</tr>
<tr>
<td>R2</td>
<td>((5 &lt; N_{MA} \leq 15) \text{ AND } (0 &lt; N_{HE} &lt; 5)) AND ((N_{NV} = 0))</td>
<td>247</td>
</tr>
<tr>
<td>R3</td>
<td>((N_{MA} \geq 15) \text{ OR } (N_{HE} &gt; 5)) OR ((N_{NV} &gt; 0))</td>
<td>254</td>
</tr>
</tbody>
</table>

\(N_{MA}, N_{HE}, N_{NV}\): number of MAs, HEs and neovessels (NV), respectively

Table 1. Retinopathy grades in Messidor dataset

4. Methods

4.1. CNN Architectures

In order to assess the strengths and limitations of CNNs, several architectures were trained and tested with particular focus on a 22 layers deep model called GoogLeNet. This very efficient network achieves state-of-the-art accuracy using a mixture of low-dimensional embeddings and heterogeneous sized spatial filters\(^2\). Increased convolution layers and improved utilization of internal network computing resources allow the network to learn deeper features. For example, the first layer might learn edges while the deepest layer learns to interpret hard exudate, a DR classification feature. The network contains convolution blocks with activation on the top layer that defines complex functional mappings between inputs and response variable, followed by batch normalization after each convolution layer. As the number of feature maps increase, one batch normalization per block is introduced in succession.

The max pooling sample-based discretization process was performed with kernel size 3x3 and stride \(^2\). The network was then flattened to one dimension after the final convolutional block. Dropout of network layers was performed until reaching the dense five node output layer, which uses a softmax activation function to compute the probability of classification labels. Leaky rectified linear unit activation was also applied with gradient value 0.01 to mitigate dead neuron bottlenecks during back-propagation\(^2\). The network uses convolutional layer L2 regularization to reduce model overfitting, cross-entropy computed error loss, and the Xavier method of initializing weights so that neuron activation functions start out in unsaturated regions.

4.2. Preprocessing

All images were converted to a hierarchical data format for preprocessing, data augmentation, and training. Preprocessing involved several steps: images were cropped using Otsu’s method to isolate the circular colored image of the retina. Images were normalized by subtracting the minimum pixel intensity from each channel and dividing by the mean pixel intensity to represent pixels in the range 0 to 1. Contrast adjustment was performed using the contrast limited adaptive histogram equalization (CLAHE) filtering algorithm.

4.3. Data Augmentation

We augmented the number of images in real-time to improve network localization capability and reduce overfitting. During each epoch, a random augmentation of images that preserve collinearity and distance ratios was performed. We implemented
random padding with zeros, zoom, rolling and rotation. These affine transformations are particularly effective when applied to disease class R1 which are the most difficult to grade and fewest in number.

4.4. Training and Testing Models

A Deep Learning GPU Training System (DIGITS) with prebuilt convolutional neural networks for image classification facilitated data management, model prototyping and real-time performance monitoring. DIGITS is an interactive system and was first used to build a classification dataset by splitting the Messidor and MildDR fundus folder into training and validation subsets of 1077 and 269 images respectively. The images were cropped to area size 256x256 and used as input data by Imagenet models previously trained for generic classification tasks. The test subset folder contained 400 images from the Lariboisiere hospital Messidor partition and was disjoint from training data. This training system, which offered extensive hyperparameter selections, was then used to build model prototypes over 100 epochs requiring approximately 20 minutes each to complete. A Tesla K80 GPU hardware device powered the training and a form of early stopping determined the optimal test set model epoch. Advanced visualization monitoring and confusion matrix statistical analysis provided important insights. The knowledge gained guided construction of more complex model architectures finely tuned for improved interpretation of our datasets see Appendix.

The more refined models were developed in Tensorflow using modified versions of the open source GitHub package (https://github.com/yidarvin/FirstAid) and evaluated with additional digital image preprocessing techniques.

4.5. Transfer Learning

Transfer learning based approaches were executed with pretrained AlexNet and GoogLeNet architectures from ImageNet. The last fully connected layer was removed, then a transfer learning scenario was followed by treating the remaining network components as a fixed feature extractor for the new dataset. The transfer learning retains initial pretrained model weights and extracts image features via a final network layer.

5. Experiments

5.1. Digital image processing improves sensitivity for mild class detection

The dataset contained images from a disparate patient population with extremely varied levels of lighting in the fundus photography. The lighting affects pixel intensity values within the images and creates unnecessary variation unrelated to classification levels. A contrast limited adaptive histogram equalization filtering algorithm, using the OpenCV (http://opencv.org/) package was applied to address this artifact. Results from this preprocessing step are visually depicted in Fig. 2. We discovered that 3-ary classifier sensitivity for the mild case increased from 0 to 29.4%, while this measure was approximately the same for the remaining two classes Fig. 3. Our digital image preprocessing technique enabled improved detection of pinpoint subtle features and microaneurysms via convolutional filters, which were previously imperceptible by the CNN. We hypothesize this change is attributable to the channel wise contrast enhancing effect of histogram equalization.

Figure 2. Contrast Limited Adaptive Histogram Equalization enhances contrast and the detection of subtle features. Shown are fundoscopic illustrations before and after CLAHE application.
5.2. Binary model classification attains benchmark performance from literature

AlexNet, VGG16 and GoogLeNet models were trained on the binary-labeled (normal or mild vs moderate to end stage) Kaggle dataset to explore strengths and weaknesses of CNNs. This implementation was performed in Tensorflow, and all model weights were allowed to be updated. The GoogLeNet model achieved the highest sensitivity of 95% and specificity of 96% using our real time data augmentation and preprocessing techniques see Fig. 4. Thus, we successfully demonstrate state-of-the-art accuracy levels that have previously been published in this field.

5.3. Multi-class training sensitivities is highly dependent on dataset fidelity

However, when we trained 3-ary and 4-ary classifiers with a GoogLeNet model on the Kaggle dataset, we were unable to achieve significant sensitivity levels for the mild class. As shown in the confusion matrix (Table 2), the sensitivity of the no DR and severe DR classes were 98% and 93% respectively; however the sensitivity for the mild class was only 7%. Thus, we find that our performance is limited by the inability of CNNs to detect very subtle features. We hypothesize that this can be explained by the considerable noise and camera artifacts as well as poor fidelity labeling (misleading or incorrect) of the Kaggle images.

Figure 4. Training Curve for model on the binary classified Kaggle data set of DR fundoscope images. Sensitivity of 95% and specificity of 96% was achieved.

<table>
<thead>
<tr>
<th></th>
<th>Pred R0</th>
<th>Pred R1</th>
<th>Pred R2 or R3</th>
</tr>
</thead>
<tbody>
<tr>
<td>True R0</td>
<td>149</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>True R1</td>
<td>21</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>True R2 or R3</td>
<td>1</td>
<td>15</td>
<td>202</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix on test set of Kaggle dataset

We gained insight into ways which might improve our results by producing visualizations that reveal at the image level our CNNs usage of high and low level features to improve probability of disease detection. Visualization of pathological...
areas shown in a heat map generated by sliding window patch-wise occlusion, Fig. 5 demonstrate a large number of false positive and false negative zones, indicating that, while large disease features and those that have a different color from the background are easily interpreted by the CNN, small and subtle features remain undetected.

Thus, we consider our Messidor collection which addresses fidelity concerns of the Kaggle dataset (see Dataset section for more details on data acquisition procedures).

After following our preprocessing steps, the 3-ary classifier achieves sensitivities for no DR and severe DR of 85% and 75% respectively as well as 29% mild class sensitivity. Given the dramatic measurement increase for the mild class, even with 5% the amount of data, it is evident that data fidelity has a strong impact on multi-class training model performance.

However, the 4-ary classifier encounters a problem of simply not having enough images to effectively train a deep CNN such as GoogLeNet. The multi-class model is unable to distinguish between different classes and behaves as a majority classifier, attempting to classify all images into a single class. In this instance overfitting is not occurring since the cross-entropy loss is nondecreasing. Rather the problem is more likely one of underfitting, where the model parameters have not been exposed to enough data requisite for discriminating between the different classes.

5.4. Transfer learning as a parallel means of exploring optimal CNN models

Our previous models relied on a Tensorflow implementation without fixed weights. The practicality of transfer learning was investigated by using a baseline prototype consisting of pretrained GoogLeNet model obtained from the ImageNet visual object recognition database. The prototype was trained on the Messidor dataset for 30 epochs using stochastic gradient descent optimization with step decay learning rate initialized at 0.002. The classification model validation achieved 66.03% as the best accuracy.

Introducing loss (and accuracy layers) to intermediate representations of the deep network allowed for faster propagation during training and avoided the vanishing gradients issue. We found that training converged much faster using the preprocessed images—in a matter of 25 epochs for the transfer learning scenario rather than 90 epochs when training on the raw data.

The accuracy measures indicate how well these intermediate inception layer representations have converged to produce the best classifier. We observed spiking behavior in the training loss curves which suggested a need for weight decay hyperparameter tuning. Additionally, tuning hyper-parameters such as L2 regularization, dropout and batch normalization produced a greater degree of accuracy layer convergence. Full training and test results can be found in the Appendix Tables 3 and 4.

However, the final test time 3-ary accuracy was similar for both, 67.2% training on the raw data and 71.25% for the transfer learning task. Grouping mild and abnormal together the accuracy was 71.5% and 74.5% respectively. The best test set accuracies can be found in Fig. 6. It is interesting to note that the accuracy of the 4-ary classifier decreases after preprocessing. We hypothesize this is due to the loss of important image features during downsampling. However, we note
that there was a gain in the overall detection of subtle features, microaneurysms, of 17% with transfer learning. This method should be further explored to determine if we can enhance sensitivity of these models to mild class DR.

6. Conclusion

Automated detection and screening offers a unique opportunity to prevent a significant proportion of vision loss in our population. In recent years, researchers have added CNNs into the set of algorithms used to screen for diabetic disease. CNNs promise to leverage the large amounts of images that have been amassed for physician interpreted screening and learn from raw pixels. The high variance and low bias of these models could allow CNNs to diagnose a wider range of nondiabetic diseases as well.

However, while we achieve state-of-the-art performance with CNNs using binary classifiers, the model performance degrades with increasing number of classes. Though it is tempting to surmise that more data may be better, previous work in the field has corroborated that CNN ability to tolerate scale variations is restricted and others have suggested that in the case of retinal images, more data cannot supplement for this inherent limitation. Gulshan et al. reported a 93-96% recall for binary classification of disease but reports that recall is not improved when training with 60,000 samples vs 120,000 samples of a private dataset.

Visualizations of the features learned by CNNs reveal that the signals used for classification reside in a portion of the image clearly visible by the observer. Moderate and severe diabetic retinal images contain macroscopic features at a scale that current CNN architectures, such as those available from the ImageNet visual database, are optimized to classify. Conversely, the features that distinguish mild vs normal disease reside in less than 1% of the total pixel volume, a level of subtleness that is often difficult for human interpreters to detect.

Medical images are fraught with subtle features that can be crucial for diagnosis. Fortunately, the most often deployed architectures have been optimized to recognize macroscopic features such as those present in the ImageNet dataset. We may therefore require a new paradigm for diagnosing diseases via CNN models. This could be a two stage lesion detection pipeline that involves feature localization followed by classification and further preprocessing steps to segment out pathologies difficult to discern by manual inspection, and finally rebalancing network weights to account for class imbalances seen in medical datasets. Overall, our future goals involve improving detection of mild disease and transitioning to more challenging and beneficial multi-grade disease detection.

7. Acknowledgements

We thank Professor Daniel Rubin MD and the National Library of Medicine for providing grant funds supporting Carson Lam and Darvin Yi, PhD Candidate in Rubin Lab, Department of Biomedical Informatics at Stanford University School of Medicine. Also acknowledged are medical students and physicians Laura C. Huang, Caroline Yu, Emma Zhao, Robert Kleinman, Ryan Smith and Ryan Shields for assistance with curation of the dataset. A related paper was presented at the ARVO annual meeting.
A. Appendix

A.1. Messidor and MildDR Raw Images

GoogLeNet Rapid Prototyping Results - Raw Images

<table>
<thead>
<tr>
<th>Model</th>
<th>Solver</th>
<th>Learning Rate</th>
<th>Policy</th>
<th>Validation Accuracy%</th>
<th>Test Set Accuracy%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-ary</td>
<td>SGD</td>
<td>1e-3</td>
<td>Step Down</td>
<td>83.82</td>
<td>72.75</td>
</tr>
<tr>
<td>2-ary</td>
<td>NAG</td>
<td>1e-3</td>
<td>Step Down</td>
<td>82.36</td>
<td>72.75</td>
</tr>
<tr>
<td>2-ary</td>
<td>Adam</td>
<td>1e-4</td>
<td>Step Down</td>
<td>86.40</td>
<td>71.75</td>
</tr>
<tr>
<td>2-ary</td>
<td>AdaGrad</td>
<td>1e-3</td>
<td>Exponential Decay</td>
<td>84.55</td>
<td>64.25</td>
</tr>
<tr>
<td>2-ary</td>
<td>RMSProp</td>
<td>1e-4</td>
<td>Sigmoid Decay</td>
<td>79.04</td>
<td>64.25</td>
</tr>
<tr>
<td>3-ary</td>
<td>RMSProp</td>
<td>1e-4</td>
<td>Exponential Decay</td>
<td>63.97</td>
<td>66.25</td>
</tr>
<tr>
<td>3-ary</td>
<td>SGD</td>
<td>1e-3</td>
<td>Step Down</td>
<td>71.69</td>
<td>64.25</td>
</tr>
<tr>
<td>3-ary</td>
<td>Adam</td>
<td>1e-4</td>
<td>Step Down</td>
<td>72.40</td>
<td>61.50</td>
</tr>
<tr>
<td>3-ary</td>
<td>NAG</td>
<td>1e-3</td>
<td>Step Down</td>
<td>69.85</td>
<td>58.75</td>
</tr>
<tr>
<td>3-ary</td>
<td>AdaGrad</td>
<td>1e-3</td>
<td>Exponential Decay</td>
<td>72.43</td>
<td>58.25</td>
</tr>
<tr>
<td>4-ary</td>
<td>Adam</td>
<td>1e-4</td>
<td>Step Down</td>
<td>67.65</td>
<td>57.25</td>
</tr>
<tr>
<td>4-ary</td>
<td>SGD</td>
<td>1e-3</td>
<td>Step Down</td>
<td>65.07</td>
<td>55.25</td>
</tr>
<tr>
<td>4-ary</td>
<td>NAG</td>
<td>1e-3</td>
<td>Step Down</td>
<td>67.71</td>
<td>58.75</td>
</tr>
<tr>
<td>4-ary</td>
<td>AdaGrad</td>
<td>1e-3</td>
<td>Exponential Decay</td>
<td>66.54</td>
<td>53.25</td>
</tr>
<tr>
<td>4-ary</td>
<td>RMSProp</td>
<td>1e-4</td>
<td>Step Down</td>
<td>66.18</td>
<td>52.75</td>
</tr>
</tbody>
</table>

Table 3. Hyperparameter optimization of the Messidor dataset trained using transfer learning on a pretrained GoogLeNet model from ImageNet. 2-ary dataset classes were group C0:R0, R1 and C1:R2, R3. 3-ary dataset classes were C0: R0, C1:R1, and C2:R2, R3. 4-ary dataset classes were C0:R0, C1:R1, C2:R2, C3:R3. C represents the label within the CNN architecture, and R represents the label from the dataset.

A.2. Messidor and MildDR Data with Data Augmented and CLAHE Images

GoogLeNet Rapid Prototyping Results - Data Augmentation, Contrast Filtering & Regularization

<table>
<thead>
<tr>
<th>Model</th>
<th>Solver</th>
<th>Learning Rate</th>
<th>Policy</th>
<th>Drop Out</th>
<th>Validati on Accuracy%</th>
<th>Test Set Accuracy%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-ary</td>
<td>Adam</td>
<td>1e-4</td>
<td>Step Down</td>
<td>(0.8,0.7,0.4)</td>
<td>88.35</td>
<td>74.50</td>
</tr>
<tr>
<td>2-ary</td>
<td>SGD</td>
<td>1e-3</td>
<td>Step Down</td>
<td>(0.6,0.6,0.4)</td>
<td>88.07</td>
<td>71.75</td>
</tr>
<tr>
<td>2-ary</td>
<td>NAG</td>
<td>1e-3</td>
<td>Step Down</td>
<td>(0.7,0.8,0.4)</td>
<td>87.50</td>
<td>70.00</td>
</tr>
<tr>
<td>2-ary</td>
<td>RMSProp</td>
<td>1e-4</td>
<td>Exponential Decay</td>
<td>(0.7,0.7,0.4)</td>
<td>85.80</td>
<td>69.50</td>
</tr>
<tr>
<td>2-ary</td>
<td>AdaGrad</td>
<td>1e-3</td>
<td>Exponential Decay</td>
<td>(0.7,0.7,0.5)</td>
<td>85.42</td>
<td>66.25</td>
</tr>
<tr>
<td>3-ary</td>
<td>AdaGrad</td>
<td>1e-3</td>
<td>Exponential Decay</td>
<td>(0.7,0.7,0.5)</td>
<td>63.28</td>
<td>68.75</td>
</tr>
<tr>
<td>3-ary</td>
<td>SGD</td>
<td>1e-3</td>
<td>Step Down</td>
<td>(0.6,0.6,0.4)</td>
<td>65.63</td>
<td>67.00</td>
</tr>
<tr>
<td>3-ary</td>
<td>NAG</td>
<td>1e-3</td>
<td>Step Down</td>
<td>(0.7,0.8,0.4)</td>
<td>67.71</td>
<td>64.75</td>
</tr>
<tr>
<td>3-ary</td>
<td>RMSProp</td>
<td>1e-4</td>
<td>Exponential Decay</td>
<td>(0.7,0.7,0.4)</td>
<td>60.68</td>
<td>64.00</td>
</tr>
<tr>
<td>3-ary</td>
<td>Adam</td>
<td>1e-4</td>
<td>Step Down</td>
<td>(0.8,0.7,0.4)</td>
<td>64.32</td>
<td>63.50</td>
</tr>
<tr>
<td>4-ary</td>
<td>SGD</td>
<td>1e-3</td>
<td>Step Down</td>
<td>(0.6,0.6,0.4)</td>
<td>60.00</td>
<td>51.25</td>
</tr>
<tr>
<td>4-ary</td>
<td>Adam</td>
<td>1e-4</td>
<td>Step Down</td>
<td>(0.8,0.7,0.4)</td>
<td>53.75</td>
<td>49.50</td>
</tr>
<tr>
<td>4-ary</td>
<td>NAG</td>
<td>1e-3</td>
<td>Step Down</td>
<td>(0.7,0.8,0.4)</td>
<td>55.00</td>
<td>47.75</td>
</tr>
<tr>
<td>4-ary</td>
<td>AdaGrad</td>
<td>1e-3</td>
<td>Exponential Decay</td>
<td>(0.7,0.7,0.5)</td>
<td>57.50</td>
<td>47.00</td>
</tr>
<tr>
<td>4-ary</td>
<td>RMSProp</td>
<td>1e-4</td>
<td>Exponential Decay</td>
<td>(0.7,0.7,0.4)</td>
<td>54.75</td>
<td>44.25</td>
</tr>
</tbody>
</table>

1 Dropout ratios for (training loss, validation loss, pool layer 5)

Table 4. Hyperparameter optimization of preprocessed Messidor dataset trained using transfer learning on a pretrained GoogLeNet model from ImageNet. 2-ary dataset classes were group C0:R0, R1 and C1:R2, R3. 3-ary dataset classes were C0: R0, C1:R1, and C2:R2, R3. 4-ary dataset classes were C0:R0, C1:R1, C2:R2, C3:R3. C represents the label within the CNN architecture, and R represents the label from the dataset.
References

Memory-Augmented Active Deep Learning for Identifying Relations Between Distant Medical Concepts in Electroencephalography Reports

Ramon Maldonado, BS, Travis R. Goodwin, MS, Sanda M. Harabagiu, PhD
University of Texas at Dallas, Richardson, TX, U.S.A.

Abstract
The automatic identification of relations between medical concepts in a large corpus of Electroencephalography (EEG) reports is an important step in the development of an EEG-specific patient cohort retrieval system as well as in the acquisition of EEG-specific knowledge from this corpus. EEG-specific relations involve medical concepts that are not typically mentioned in the same sentence or even the same section of a report, thus requiring extraction techniques that can handle such long-distance dependencies. To address this challenge, we present a novel framework which combines the advantages of a deep learning framework employing Dynamic Relational Memory (DRM) with active learning. While DRM enables the prediction of long-distance relations, active learning provides a mechanism for accurately identifying relations with minimal training data, obtaining an 5-fold cross validation F1 score of 0.7475 on a set of 140 EEG reports selected with active learning. The results obtained with our novel framework show great promise.

Introduction
Clinical electroencephalography (EEG) is the most important investigation in the diagnosis and management of epilepsies and other types of brain disorders1. An EEG records signals measured along the scalp, which can be correlated with brain activity, enabling the diagnosis of brain-related illnesses. However, the complexity of EEG signals complicates its interpretation, usually documented in EEG reports. With the advent of big collections of clinical EEGs, the interpretation of EEG signals can be improved by providing neurologists with results of search for patients that exhibit similar EEG characteristics. Recently, Goodwin & Harabagiu (2016)2 have described the MERCuRY (Multi-modal EncephalogRam patient Cohort discoveRY) system that relies on deep learning to represent the EEG signal and operates on a multi-modal EEG index resulting from the automatic processing of both the EEG signal and the EEG reports. The MERCuRY system allows neurologist to search a vast data archive of clinical electroencephalography (EEG) signals and EEG reports, enabling them to discover patient populations relevant to queries like Q: Patients with shifting arrhythmic delta suspected of underlying cerebrovascular disease?

The identification of relevant patient cohorts satisfying the characteristics expressed in queries such as Q relies on the ability to automatically and accurately recognize both in the queries and throughout the EEG reports (a) various clinical concepts and their attributes as well as (b) relevant relations between them. An active deep learning paradigm capable of accurately identifying medical concepts such as (concept 1): "shifting arrhythmic delta", an EEG activity, and (concept 2): "cerebrovascular disease", a medical problem, mentioned in Q, was reported in Maldonado et al. (2017)3. The same system identifies EEG activities and events as well as medical problems and treatments in EEG reports of patients from the desired cohort, e.g.:

MEDICATIONS: [Pantoprazole]TREATMENT, [Folic Acid]TREATMENT, [Carvedilol]TREATMENT
INTRODUCTION: Digital video EEG was performed at the bedside using standard 10-20 system of electrode placement with 1 channel EKG.
DESCRIPTION OF THE RECORD: The background EEG is characterized by [slowing]EEG_ACTIVITY and [disorganization]EEG_ACTIVITY. There is prominent shifting arrhythmic [delta activity]EEG_ACTIVITY more prominent in the left mid to anterior temporal region. [Photic stimulation]EEG_EVENT generates scant [driving]EEG_ACTIVITY.
IMPRESSION: Abnormal EEG due to:
1. Marked background [slowing]EEG_ACTIVITY and [disorganization]EEG_ACTIVITY
2. Some arrhythmic [delta activity]$^{EEG,ACTIVITY}$

CLINICAL CORRELATION: These findings are supportive of a [bihemispheric disturbance of cerebral function]$^{MEDICAL,PROBLEM}$

These are nonspecific findings which can be seen in a toxic and metabolic [encephalopathy]$^{MEDICAL,PROBLEM}$ and/or underlying [cerebrovascular disease]$^{MEDICAL,PROBLEM}$. In order to identify a relevant patient, his or her EEG reports need to be relevant to the query. The identification of the medical concepts from the query in the EEG report is not sufficient, as many false positives can be produced. For example, not all patients having shifting arrhythmic delta activity (medical concept 1 from the Query) can be suspected of underlying cerebrovascular disease (concept 2 from the Query), unless a relation between these two medical concepts can be inferred from the EEG report. But, as shown in our EEG report example, mentions of the two concepts from the query $Q$ in the EEG report, underlined and indexed with their respective concept numbers, do not appear in the same sentence, or even in the same section of the EEG report. Therefore, current state-of-the-art methods of identifying relations between medical problems$^{4–6}$ cannot be used, as they expect the arguments of a relation to be present in the same sentence of a clinical document. Consequently, we needed to develop a novel method for identifying relations between medical concepts automatically recognized by the method reported in Maldonado et al. (2017)$^{3}$. The relation identification approach reported in this paper operates on pairs of concepts from the same EEG report, that are not constrained to appear in the same sentence or section of the report. Our method identifies in the exemplified EEG report seven relations between medical concepts*: (R1) [delta activity]$^{EEG,ACTIVITY}$ – [cerebrovascular disease]$^{MEDICAL,PROBLEM}$; (R2) [slowing]$^{EEG,ACTIVITY}$ – [bihemispheric disturbance of cerebral function]$^{MEDICAL,PROBLEM}$; (R3) [disorganization]$^{EEG,ACTIVITY}$ – [bihemispheric disturbance of cerebral function]$^{MEDICAL,PROBLEM}$; (R4) [bihemispheric disturbance of cerebral function]$^{MEDICAL,PROBLEM}$ – [encephalopathy]$^{MEDICAL,PROBLEM}$; (R5) [Pantoprazole]$^{MEDICAL,PROBLEM}$ – [TREATMENT–TREATMENT-FOR] – [GI bleed]$^{MEDICAL,PROBLEM}$; (R6) [Folic acid]$^{TREATMENT}$ – [TREATMENT–TREATMENT-FOR] – [anemia]$^{MEDICAL,PROBLEM}$; (R7) [photic stimulation]$^{EEG,EVENT}$ – [driving]$^{EEG,ACTIVITY}$. The relation R1 warrants the relevance of the exemplified EEG report to the query $Q$, identifying a patient from the desired cohort. When relations between medical concepts are recognized automatically in a large collection of EEG reports, they enable the generation of EEG-specific knowledge embeddings of high accuracy. High-quality embeddings have been shown recently$^{7}$ to be crucial in designing relevance models that rely on deep learning, and thus produce excellent results.

Background

In recent work$^{8}$, we have proposed a novel paradigm for learning knowledge from a large corpus of EEG reports enabled by deep learning methods that observe the likelihood that medical concepts share certain relations, as evidenced by data. The resulting knowledge representation, called medical knowledge embeddings (MKE), resolves the semantic heterogeneity which arises when different terminology is used to refer to the same concepts or relations. For example, noted in Sahoo et al. (2014)$^{9}$, a seizure with alteration of consciousness may be referred to as complex partial seizure, dialectic seizure, or focal dyscognitive seizure by different epilepsy experts. An MKE representation should place all these expressions in a similar location of the multi-dimensional space, as it learns that they are involved in similar relations with other epilepsy-relevant concepts. When analyzing the results of the MKE produced from the EEG reports, it became clear that the quality of the identified relations needed to be improved$^{8}$. Instead of considering all potential relations, only accurate relations between medical concepts should be used in the MKE.

Although it is now well established that automated extraction of knowledge from clinical notes involves accurately identifying not only the medical concepts, but also the various relationships in which they are involved$^{10}$, the automatic identification of relations between medical concepts in EEG reports is hindered by two major obstacles. First, most of the successful techniques for automatically recognizing concepts in clinical texts considered only the target relations from the 2010 i2b2/VA challenge$^{11}$, which include relations from the following 3 categories: medical problem—treatment (TrP) relations, medical problem—test (TeP) relations, and medical problem—medical problem (PP) relations. As illustrated in the exemplified EEG report, TrP relations would be useful, but other types of relations relevant for the knowledge expressed in the report would be missed. The second hurdle is generated by the constraint that only relations between medical concepts observed in the same sentence can be identified with current methods, even those using deep learning methods capable of processing large corpora of clinical documents$^{6}$. We found the so-

---

*The definitions of the relations is provided in the annotation schema detailed in the Methods section.*
olution of both these limitations by considering and extending RelNet\textsuperscript{12}, a memory-augmented neural network in which medical concepts can be processed in abstract memory cells while relations between medical concepts are processed in separate relation memory cells. The memories implicitly model the current knowledge about medical concepts and the relations they share.

**Data**

In this work, relations between medical concepts were discovered in the EEG reports publicly available from the Temple University Hospital (TUH), comprising over 25,000 EEG reports from over 15,000 patients collected over 12 years. EEG reports are designed to convey a written impression of the visual analysis of the EEG along with an interpretation of its clinical significance. In accordance with the American Clinical Neurophysiology Society Guidelines for writing EEG reports, the reports from the TUH EEG Corpus start with a clinical history of the patient including information about the patient’s age, gender, current medical conditions (e.g. “change in mental status”), and relevant past medical conditions (e.g. “history of GI bleed”) followed by a list of medications the patient is currently taking (e.g. “Pantoprazole”). Together, these two initial sections depict the clinical picture of the patient, containing a wealth of medical concepts including medical problems (e.g. “encephalopathy”), symptoms (e.g. “without heart rate”), signs (e.g. “triching” and treatments (e.g. “Depakote”, “pacemaker”). After the clinical picture of the patient is established, the introduction section of the EEG report describes the techniques used for the current EEG (e.g. “digital video EEG using standard 10-20 system of electrode placement with one channel of EKG”), the patient’s condition at the time of the record (e.g. fasting, asleep), and possible activating procedures carried out (e.g. “photic stimulation”).

The description section is the mandatory part of the report, meant to provide a complete and objective description of the EEG, noting all observed EEG activities (e.g. “delta activity”), patterns (e.g. “slowing”), and EEG events (e.g. “photic stimulation”, “myoclonus”). The impression section indicates whether or not the EEG test is abnormal and, if so, lists the abnormalities in decreasing order of importance. These abnormalities are usually characteristic EEG activities (e.g. “arrhythmic delta activity”), but can also be EEG Events (e.g. “clinical seizures characterized by myoclonus”). Finally, the clinical correlation section explains what the EEG findings mean in terms of clinical interpretation, (e.g. “findings are supportive of bihemispheric disturbance of cerebral function”).

**Preprocessing.** To identify medical concepts in each EEG report automatically, we relied on the active deep-learning system described in Maldonado et al. (2017)\textsuperscript{3}, capable of accurately recognizing EEG Activities, EEG Events, medical problems, treatments, and tests. The International Federation of Clinical Neurophysiology defines an EEG activity as “any EEG wave or sequence of waves”, and an EEG event as “any stimulus that activates the record”\textsuperscript{13}. For each of these medical concepts, a set of attributes is also recognized. For EEG activities we identified (a) general attributes of the waves, e.g. the MORPHOLOGY, the FREQUENCY BAND or the MAGNITUDE; (b) temporal attributes, e.g. RECURRENCE and (c) spatial attributes, e.g. DISPERSAL, HEMISPHERE and BRAIN LOCATION. As reported in Maldonado et al. (2017)\textsuperscript{3} we recognize 18 different attributes for each EEG Activity. Two of these attributes are the modality and the polarity of the EEG activity. When we considered the recognition of the modality, we took advantage of the definitions used in the 2012 i2b2 challenge on evaluating temporal relations in clinical text\textsuperscript{14}. In that challenge, modality was used to capture whether a clinical event discerned from a medical record actually happens, is merely proposed, mentioned as conditional, or described as possible. We extended this definition such that the possible modality values of factual, possible, and proposed indicate that clinical concepts mentioned in the EEGs are actual findings, possible findings and findings that may be true at some point in the future, respectively. For identifying polarity of clinical concepts in EEG reports, we relied on the same definition used in the 2012 i2b2 challenge, considering that each concept can have either a positive or a negative polarity, depending on any absent or present negation of its finding. Through the identification of modality and polarity of the clinical concepts, we aimed to capture the neurologists beliefs about the clinical concepts mentioned in the EEG report. For example, the EEG activity “high amplitude spike and slow wave complexes seen anteriorly on the left” can be more completely described by noting its location (head region), hemisphere (side of the brain), magnitude (describes the activity’s amplitude), and morphology (the type or form of the EEG wave). The full attribute specification is described in Maldonado et al. (2017)\textsuperscript{3}.

In addition, we also automatically identified EEG events, medical problems, treatments, and tests, which are characterized only by two attributes: modality and polarity. When pre-processing the corpus of EEG reports using the system
developed in Maldonado et al. (2017)\(^3\), we were able to detect 365,218 medical concept mentions with 3,062,846 attributes in the TUH EEG corpus. For the remainder of this paper, the term concept will refer to a medical concept and its attributes. Moreover, because mentions of the same medical concept may be expressed with different words (e.g. “epileptic attack” and “epileptic seizure” are both mentions of the medical problem [epilepsy]\(\text{MEDICAL\_PROBLEM}\)) we normalized each concept mention into a canonical form using the morphology attribute for EEG activities and the preferred name of each medical problem and treatment given by the Unified Medical Language System (UMLS)\(^15\). EEG events are normalized into a set of 10 events types using the same heuristic approach as in our previous work\(^8\). In this work, we use a subset of 140 EEG reports selected via active learning (explained in the Methods section) to train and evaluate our system for relation identification.

Methods

The automatic identification of relations between pairs of automatically identified medical concepts in EEG reports, regardless of their presence in the same sentence, section or across sentences and sections of the report, has been made possible by a novel deep learning system that we designed and implemented, namely the Memory-Augmented Active Deep Learning (MAADL) system. MAADL combines the strength of the Active Learning framework with the advantages of deep learning. While deep learning methods provide unprecedented performance in many tasks, active learning allows a deep learner to achieve this performance with less manually annotated training data, as it exposes the system to new examples on which its performance is still suffering. The operation of MAADL, which is illustrated in Figure 1, relies on the availability of automatically recognized medical concepts in the corpus of EEG reports obtained with the deep-learning approach reported in Maldonado et al. (2017)\(^3\). The identification of relations between medical concepts in MAADL uses the following five steps:

**STEP 1**: The development of an annotation schema for relations between medical concepts in EEG reports;

**STEP 2**: Annotation of relations between medical concepts in the initial training data;

**STEP 3**: Design of a deep learning method for detecting relations between medical concepts in the EEG reports;

**STEP 4**: Development of sampling methods for the MAADL;

**STEP 5**: Usage of the Active Learning system which involves:

**STEP 5.a**: Accepting/Editing annotations of sampled examples of relations between medical concepts in EEG reports;

**STEP 5.b**: Re-training the deep learning method and evaluating the re-trained system.

---

**Figure 1**: The Memory-Augmented Active Deep Learning (MAADL) system for automatically identifying relations between pairs of medical concepts in EEG reports.
STEP 1: Annotation Schema for relations between pairs of medical concepts in EEG reports: The annotation of relations benefits from the definitions of relations developed in our previous work\(^3\). First, we decided to consider only binary relations between four types of medical concepts: (1) EEG events; (2) EEG activities; (3) medical problems and (4) treatments. Second, we decided to consider only three types of relations between medical concepts: EVIDENCES, EVOKES, and TREATMENT-FOR. The EVIDENCES relation considers (a) EEG events, EEG activities, treatments, and medical problems as providing evidence for (b) medical problems mentioned in the EEG report. The EVOKES relation represents the relationship where a medical concept evokes an EEG activity. EEG events, other EEG activities, medical problems and treatments can all evoke EEG activities. The TREATMENT-FOR relation links treatments to the medical problems for which they are prescribed. In addition, we made the decision to annotate relations between medical concepts, and not between their mentions in the EEG report. Because the same concept can be mentioned multiple times in the same EEG report, the representation of concepts achieved while pre-processing the EEG reports by (i) their normalized mention and (ii) their attributes made it possible to recognize co-referring mentions of the same concept by simply grouping concepts with the same normalized mention name and attribute values. Therefore, all co-referring mentions were considered a unique concept, and relations were annotated between unique concept pairs.

STEP 2: Initial Relation Annotations: A set of 40 EEG reports with 198 EVIDENCES relations, 146 EVOKES relations, and 72 TREATMENT-FOR relations were manually annotated and used as the initial training data for the relation detection system. This set of reports had previously been manually annotated with medical concepts and their attributes to ensure errors in concept/attribute detection did not affect relation detection.

STEP 3: Design of Deep Learning Architecture for the Memory-Augmented Active Deep Learning System: We designed a deep learning architecture, called EEG-RelNet, which provides an end-to-end detection of relations between medical concepts in each EEG report by using a neural network augmented with two types of memories: (i) a memory for each medical concept; and (ii) a memory for each relation between each pair of medical concepts. Moreover, the relational memory is dynamic as it changes to model the specific relations observed in each EEG report.

STEP 4: Active Learning Sampling Method: To improve the quality of the identified relations between medical concepts in EEG reports, as illustrated in Figure 1, an active learning loop is designed. In an active learning framework, the sampling method is used to automatically select examples of relations for human validation. Since this work is focused on relation detection between pairs of medical concepts, we chose a sampling method that only prioritizes relation detection performance, ignoring the quality of medical concepts and their attributes. Therefore, we do not use the rank combination protocol reported in our previous work\(^3\), opting for standard uncertainty sampling\(^6\) whereby EEG reports containing relations for which the model is most uncertain are selected for manual validation. The uncertainty of a report is measured at the report level by averaging the uncertainty of each relation classification decision in the report. The uncertainty of a relation classification decision is calculated using Shannon Entropy, \(H(R) = - \sum_i R_i \log R_i\), where \(R\) is a vector representing the probability distribution over possible relation types. These probability distributions are derived by EEG-RelNet from the learned dynamic relation memory, as shown in Figure 1.

STEP 5: Usage of the Memory-Augmented Active Deep Learning System: As shown in Figure 1, each iteration of active learning involves using the EEG-RelNet to make automatic relation annotations on the unlabeled EEG reports, selecting the most informative examples for manual validation, and re-training the EEG-RelNet using the new set of validated training examples.

EEG-RelNet: a Deep Learning Architecture for long-distance Relation Detection in EEG Reports

While medical concepts (EEG activities, EEG events, medical problems, treatments and tests) are available in each EEG report, due to the preprocessing that was applied to the entire TUH EEG corpus, inference of the EVIDENCES, EVOKES, and TREATMENT-FOR relations between pairs of such concepts was produced through dynamic memories based on neural networks, capable to capture the implicit participation of each medical concept in a relation of interest. Inspired by RelNet, a model reported in Bansal et al. (2017)\(^2\), we developed the EEG-RelNet, a deep neural network architecture that operates on the full text of an EEG report considering all medical concepts identified in the report to detect relations of the type EVIDENCES, EVOKES, and TREATMENT-FOR between any pair of concepts. More specifically, given the full text of an EEG report and the set of medical concepts identified in that report, EEG-RelNet can predict whether there is relation of type \(t\), \(R_{ij}^t\), between any pair of medical concepts \(c_i\) and \(c_j\) recognized in the report. To do so, EEG-RelNet processes the EEG report, one sentence at a time, reading its words, encoding the
information from the sentence, processing the sentence information in the dynamic relational memory, and predicting each type of relation based on the dynamic memories after they have processed each sentence in the EEG report. The three modules of EEG-RelNet are:

- the **Input Encoding Module** which encodes information from the report at concept- and sentence-level embedding vectors, which are used throughout the deep learning architecture;
- the **Dynamic Relational Memory Module** which maintains and updates a set of hidden states called memories to capture accumulated information about each medical concept and potential relation in the EEG report;
- the **Output Module** which uses the updated memories to determine the most likely relations (and their types) between medical concepts in the EEG report.

In the remainder of this section, we provide a detailed description of each module of EEG-RelNet.

**The Input Encoding Module.** The role of this module is to learn (1) an embedding encoding each medical concept as well as each of its attributes and (2) an embedding encoding the information from each sentence in the EEG report. Formally, we represent an EEG report as a set of medical concepts, \( C = \{c_1, \cdots, c_d\} \), and a sequence of sentences, \([s_1, \cdots, s_n]\). Each medical concept, \( c_i \), is associated with several \( N \)-dimensional vectors called embeddings: (a) an embedding for the normalized concept name, \( c_i \in \mathbb{R}^N \) and (b) separate embeddings for each of its \( A \) attributes values \( \{a_i^1, \cdots, a_i^A\} \). Thus, the embedding \( c_i \) for a medical concept is created by (1) concatenating the embedding for the name of the medical concept with the embedding for each of its attributes and (2) projecting this concatenated vector using a learned weight matrix \( W_C \in \mathbb{R}^{N \times (|A| + 1)} \), i.e. \( c_i = W_C \times [c_i^1, a_i^1, \cdots, a_i^A] \). In this way, each medical concept is represented by an embedding, \( \tilde{c}_i \), which is a vector in \( \mathbb{R}^N \).

Participation of medical concepts in relations is informed by the context of each concept in the text of the EEG report. Contextual information is provided by the words of the sentence where the concept is mentioned, hence a representation of words from each sentence as is also desirable. Therefore, we learn an embedding \( e_i \) for each word \( u_i \) in a sentence, enabling us to represent each sentence as a sequence of embeddings \( E = [e_1, \cdots, e_m] \) such that the elements of \( E \) occur in the same order as the words from the sentence \(^1\). While the traditional choice for combining and composing the embeddings in \( E \) into a single sentence embedding would be a Recurrent Neural Network (RNN), we instead adopt a more recent and significantly more efficient strategy, namely a positional mask\(^12,17\) such that the \( k \)-th sentence from the EEG report is represented as: \( s_k = \sum_i f_i \odot e_i \), given that the sentence had \( m \) words, and the vectors \( [f_1, \cdots, f_m] \), represent the learned positional mask and \( \odot \) is the element-wise product. It is important to note that the same vectors \( [f_1, \cdots, f_m] \) are used when each new sentence is encoded and they are learned jointly with the other parameters of the deep learning model.

**The Dynamic Relational Memory Module.** Because EEG reports often contain long-distance relations between concepts we relied on a Dynamic Relational Memory\(^12\) (DRM) Module to keep track of the interactions between medical concepts in each report. The DRM accumulates information about medical concepts and the relations between them by processing each sentence encoded by the Input Module and updating a set of hidden states, called memories. Specifically, given a sentence embedding, \( s_k \) and the corresponding set of concept embeddings \( \{\tilde{c}_1, \cdots, \tilde{c}_d\} \), there are two scenarios for each \( \tilde{c}_i \): (scenario 1): the medical concept \( c_i \) has not been mentioned in any previous sentence, thus its Concept Memory needs to be accounted for using a single, shared Concept Memory Cell; and (scenario 2): the concept \( c_i \) has been previously mentioned, and thus its corresponding Concept Memory needs to be updated. Moreover, since each medical concept \( c_i \) may participate in a relation, in (scenario 1), a unique Relation Memory needs to account for each relation in which the concept participates, whereas in (scenario 2) the corresponding Relation Memory needs to be updated. If an EEG report refers to \( d \) medical concepts, there will be \( d \times (d - 1) \) Relation Memory Cells. The **Dynamic Relational Memory (DRM)** consists of the entire set of Concept and Relation Memories in an EEG report.

The Concept Memories are organized as a Key-Value Memory Network\(^18\). Key-value paired memories are a generalization of the way context of concepts is stored in memory. In a Key-Value Memory Network, the lookup (addressing) stage is based on the key vector while the reading stage (giving the returned result) returns the value memory. Consequently, in EEG-RelNet, memory vectors are tied to so-called key vectors enabling the model to only update a memory

---

\(^1\)Embeddings, \( e_{u_i} \), corresponding to words contained within a concept mention, \( c_i \), are replaced with the embedding for that concept instead of the word, i.e. \( e_{u_i} = \tilde{c}_i \). This is required to enable the Key-Value memory structures described in the next subsection.
As illustrated in Figure 2, when each sentence \( s_i \) is processed, the DRM uses and updates not only concept memories, but also a much larger set of relation memories. This is explained by the fact that unfortunately, maintaining a single memory vector for each concept is not sufficient for modeling concepts that participate in multiple relations, especially when those relations involve concepts that are mentioned at significant distance in the EEG report. Thus, to model memory vectors when the input sentence has context that is relevant to medical concept \( c_i \), given a medical concept embedding, \( c_i \), and a sentence encoding, \( \vec{s}_k \), via the following equations:

\[
\begin{align*}
g^c_i &= \sigma \left( \langle \vec{s}_k, h_i + c_i \rangle \right) \\
\tilde{h}_i &= \phi \left( W_u h_i + W_v c_i + W_s \vec{s}_k \right) \\
h_i &\leftarrow h_i + g^c_i \odot \tilde{h}_i
\end{align*}
\]

where \( W_u, W_v \) and \( W_s \) are trainable weight matrices in \( \mathbb{R}^{N \times N} \), \( \langle \cdot, \cdot \rangle \) is the inner product, \( \sigma \) is the sigmoid function and \( \phi \) is a Parametric Rectified Linear Unit (PReLU). Equation 1 is a gating function that determines how much the \( k^{th} \) input sentence affects the \( i^{th} \) Concept Memory such that \( g^c_i \in [0, 1] \) values close to 1 indicate sentence \( s_k \) relevant to medical concept \( c_i \) and values close to 0 indicate the opposite. Equation 2 defines the candidate Concept Memory that will be used to update the existing Concept Memory, \( h_i \), after it is scaled by \( g^c_i \) as shown in equation 3.

As illustrated in Figure 2, when each sentence \( s_i \) is processed, the DRM uses and updates not only concept memories, but also a much larger set of relation memories. This is explained by the fact that unfortunately, maintaining a single memory vector for each concept is not sufficient for modeling concepts that participate in multiple relations, especially when those relations involve concepts that are mentioned at significant distance in the EEG report. Thus, to model interactions each concept has with each other concept in the same EEG report, we maintain a set of Relation Memories for each sentence.

Figure 4: The Dynamic Relational Memory Module of EEG-RelNet. The Dynamic Relational Memory Module processes \( n \) sentences, updating a set of \( d \) Concept Memories and \( d (d - 1) \) Relation Memories for each sentence.

where \( W_A \) and \( W_B \) are trainable weight matrices in \( \mathbb{R}^{N \times N} \). As in the Concept Memory Cell, the Relation Memory

\[
\begin{align*}
g^r_{ij} &= g^c_i g^c_j \sigma \left( \langle \vec{s}_k, r_{ij} \rangle \right) \\
\tilde{r}_{ij} &= \phi \left( W_A r_{ij} + W_B \vec{s}_k \right) \\
r_{ij} &\leftarrow r_{ij} + g^r_{ij} \odot \tilde{r}_{ij}
\end{align*}
\]
Cell uses a gating function (equation 4) and a candidate memory (equation 5) to update the Relation Memory in a way that reflects how relevant the input sentence, \(s_k\), is to the concept pair, \((c_i, c_j)\). To compute the gate value \(g^j_i\), the Relation Memory Cell uses the two concept gate values, \(g^j_i, g^j_i\) from the Concept Memory Cells for concepts \(c_i\) and \(c_j\), ensuring that input sentences that are relevant to either concept can be used to update the Relation Memory. By maintaining a memory vector for each pair of concepts and updating that memory vector as the model accumulates information across each sentence in an EEG report, EEG-RelNet can be interpreted as constructing a local latent knowledge graph\(^{12}\) for each EEG report, where each Relation Memory represents a possible relation in the graph.

The Output Module. The output module makes use of the Dynamic Relational Memory updated after processing the last sentence in the EEG report to identify relations (and their types) between any pair of medical concepts from the report. The relation prediction, \(\hat{R}_{ij}\) between medical concepts \(c_i\) and \(c_j\) is produced by passing the Concept Memories associated with concepts \(c_i\) and \(c_j\) along with the Relational Memory \(r_{ij}\) to two fully connected PReLU layers followed by a softmax layer: where \(W_q \in \mathbb{R}^{N \times 3N}\) and \(W_z \in \mathbb{R}^{1 \times N}\) are learned weight matrices, and \(\phi\) is

\[
q_{ij} = \phi(W_q [h_i, h_j, r_{ij}])
\]

(7)

\[
R^j = \text{softmax}(\phi(W_z q_{ij}))
\]

(8)
a Parametric Rectified Linear Unit. \(R^j\) is a probability distribution over 4 possible relations: the 3 relation types described in the annotation schema and a 4\(^{th}\) type indicating no relation. Consequently, the relation (if any) detected between concepts \(c_i\) and \(c_j\) is given by \(\hat{R}_{ij} = \text{argmax}_r R^j_r\).

Results

To evaluate the Memory-Augmented Active Deep Learning (MAADL) system, we measured (1) the performance of EEG-RelNet and (2) the impact of Active Learning. To measure the impact of the EEG-RelNet architecture, we compare our system with two alternate configurations and one baseline:

1. **EEG-RelNet\(_{\text{NRM}}\)** is a deep neural network structured similarly to EEG-RelNet but without Relation Memories. Formally, we replace equation 7 with \(q_{ij} = \phi(W_q [h_i, h_j])\), and equations [4-6] are not used.

2. **EEG-RelNet\(_{\text{NA}}\)** is a deep neural network structured similarly to EEG-RelNet that ignores the attributes of each medical concept in the Input Module. Formally, EEG-RelNet\(_{\text{no-atrr}}\) represents each concept embedding using only the embedding for the name of that concept, \(\vec{c}_i = \vec{c}_i\).

3. **Heuristic** is a simple rule-based baseline from Maldonado et al. (2017)\(^8\) that uses medical concept type and section type to detect relations. EVIDENCES relations are created between any medical concept in an EEG report and medical problems in the clinical correlation section, EVOKES relations are created between any medical concept and an EEG activity, and TREATMENT-FOR relations are created between any treatment and medical problems in the history section of the EEG report.

We follow the evaluation procedure for relation classification reported in the 2012 Informatics for Integrating Biology at the Bedside (i2b2) shared task\(^{14}\). The Precision, Recall, and \(F_1\) measure for each relation type are calculated using 5-fold cross validation on the full set of 140 manually annotated EEG reports containing 1513 relations between 3691 medical concepts. Each EEG-RelNet configuration is trained for 10 epochs with the same random initialization, using \(N=100\) as the embedding size.

EEG-RelNet is able to successfully detect the three relation types, EVOKES, EVIDENCES, and TREATMENT-FOR, obtaining \(F_1\) scores of 0.8371, 0.6939, and 0.7116, respectively. Clearly, EEG-RelNet obtains the best performance on each relation type, demonstrating the importance of both the Dynamic Relational Memory and medical concept attributes when detecting relations. EEG-RelNet achieves significantly better performance when recognizing EVOKES relations compared to the other two relation types indicating that the network is able to correctly link medical problems with the EEG activities they evoke. The effect of the Dynamic Relational Memory is most obvious when considering the EVIDENCES relation type, increasing the \(F_1\) measure by nearly 20\%. Interestingly, the removal of attribute information from the model drastically reduces performance when detecting the EVOKES relation type, but only slightly reduces performance on the other two types compared to the EEG-RelNet\(_{\text{NRM}}\) system. The Heuristic approach is able to achieve the highest recall on each relation type since it was specifically designed for high recall. However, due to the poor precision, the Heuristic baseline achieves by far the worst overall performance.

To evaluate the performance of Memory-Augmented Active Deep Learning (MAADL), we measured the change in
performance after each additional round of active learning. Figure 5 presents these results, clearly showing a significant increase in performance from a macro-average F1 score of 0.6040 to 0.7475 (23.75%) with only 100 additional EEG reports annotated.

### Discussion

In general, EEG-RelNet is able to correctly recognize relations between medical concepts as indicated by the macro-average F1 score of 0.7475. However, EEG-RelNet is clearly able to recognize EVOKE relations more accurately than EVIDENCE and EVIDENCES relations, with F1 scores of 0.8371, 0.7116, and 0.6939, respectively. We believe the superior performance when detecting EVOKE relations may be explained by (1) the fact that EVOKE relations always involve an EEG activity and (2) the more sophisticated representation of EEG activities compared to the other medical concepts. Specifically, EEG activities have 18 semantic attributes that capture rich information, but medical problems, treatments, and EEG events only have two attributes: modality and polarity. This suggest that semantic attributes play an important role in detecting relations between medical concepts. We believe the performance of our model could be improved in the future by introducing more sophisticated representations of medical problems, treatments, and EEG events using neurological ontologies or other sources of medical knowledge, like the Unified Medical Language System (UMLS). For example, when determining if the concept [Lamictal] is a TREATMENT FOR the concept [seizure], it would be beneficial to know that Lamictal is an anticonvulsant - knowledge contained in the UMLS. Another interesting phenomenon revealed by our experiments is that the recall when detecting TREATMENT-FOR relations is especially high (0.8953) while the precision is low (0.5905). A possible explanation for this phenomenon is that medications (the most common type of treatment in an EEG report) are listed contiguously in the list of medications section. Consequently, the model has difficulty determining which treatments from the same list are TREATMENT-FOR specific medical problems. This kind of error is especially prevalent at the beginning of active learning, but we can see from Figure 5 that performance sharply increases between annotation rounds 4-6, as the model is introduced to more TREATMENT-FOR relations. As described in the Background, in previous work we demonstrated that the types of relations detected by the MAADL system can be used to capture domain-specific medical knowledge in the form of Medical Knowledge Embeddings (MKE). However, one of the main limitations of the MKE was the poor quality of the automatically detected relations. Using MAADL to more accurately detect relations should enable higher quality MKE to be learned.

<table>
<thead>
<tr>
<th>Metric</th>
<th>EEG-RelNet</th>
<th>EEG-RelNet_NA</th>
<th>EEG-RelNet_NA</th>
<th>Heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EVOKES</strong></td>
<td>Precision</td>
<td>0.8563</td>
<td>0.8037</td>
<td>0.6605</td>
</tr>
<tr>
<td>Recall</td>
<td>0.8187</td>
<td>0.7500</td>
<td>0.6088</td>
<td><strong>0.9771</strong></td>
</tr>
<tr>
<td>F1</td>
<td><strong>0.8371</strong></td>
<td>0.7759</td>
<td>0.6336</td>
<td>0.3265</td>
</tr>
<tr>
<td><strong>EVIDENCES</strong></td>
<td>Precision</td>
<td><strong>0.7086</strong></td>
<td>0.6325</td>
<td>0.6193</td>
</tr>
<tr>
<td>Recall</td>
<td>0.6798</td>
<td>0.5365</td>
<td>0.5506</td>
<td><strong>0.8624</strong></td>
</tr>
<tr>
<td>F1</td>
<td>0.6939</td>
<td>0.5805</td>
<td>0.5829</td>
<td>0.2910</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>EEG-RelNet</th>
<th>EEG-RelNet_NA</th>
<th>EEG-RelNet_NA</th>
<th>Heuristic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TREATMENT-FOR</strong></td>
<td>Precision</td>
<td>0.5905</td>
<td><strong>0.5932</strong></td>
<td>0.5268</td>
</tr>
<tr>
<td>Recall</td>
<td>0.8953</td>
<td>0.8845</td>
<td>0.8520</td>
<td><strong>0.9856</strong></td>
</tr>
<tr>
<td>F1</td>
<td><strong>0.7116</strong></td>
<td>0.7101</td>
<td>0.6510</td>
<td>0.2921</td>
</tr>
<tr>
<td><strong>All Relations (Macro Average)</strong></td>
<td>Precision</td>
<td><strong>0.7185</strong></td>
<td>0.6764</td>
<td>0.6022</td>
</tr>
<tr>
<td>Recall</td>
<td>0.7979</td>
<td>0.7237</td>
<td>0.6705</td>
<td><strong>0.9417</strong></td>
</tr>
<tr>
<td>F1</td>
<td><strong>0.7475</strong></td>
<td>0.6888</td>
<td>0.6225</td>
<td>0.3032</td>
</tr>
</tbody>
</table>

Table 1: Performance of the MAADL system for detecting relations between medical concepts in EEG reports.
Conclusion

In this paper we describe a novel active deep learning framework for identifying relations between medical concepts discussed in the text of EEG reports by making use of the EEG-RelNet, a neural architecture capable of inferring relations between concepts through its dynamic relational memory. This deep learning architecture allowed us to identify relevant relations between medical concepts that were not mentioned in the same sentence or section of the EEG report, which is a key contribution of the framework presented in this paper. Most previous methods successfully employed for recognizing relations between medical concepts from clinical documents addressed only the case when the concepts were observed in the same sentence.

Acknowledgements

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References


Accurate and interpretable intensive care risk adjustment for fused clinical data with generalized additive models

Ben J. Marafino, BS1, R. Adams Dudley, MD, MBA2, Nigam H. Shah MBBS, PhD3, Jonathan H. Chen, MD, PhD3

1Biomedical Informatics Training Program, Stanford University, Stanford, CA; 2Center for Healthcare Value, Philip R. Lee Institute for Health Policy Studies, and Department of Pulmonary and Critical Care Medicine, University of California, San Francisco, San Francisco, CA; 3Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA

Abstract

Risk adjustment models for intensive care outcomes have yet to realize the full potential of data unlocked by the increasing adoption of EHRs. In particular, they fail to fully leverage the information present in longitudinal, structured clinical data – including laboratory test results and vital signs – nor can they infer patient state from unstructured clinical narratives without lengthy manual abstraction. A fully electronic ICU risk model fusing these two types of data sources may yield improved accuracy and more personalized risk estimates, and in obviating manual abstraction, could also be used for real-time decision-making. As a first step towards fully ‘electronic’ ICU models based on fused data, we present results of generalized additive modeling applied to a sample of over 36,000 ICU patients. Our approach outperforms those based on the SAPS and OASIS systems (AUC: 0.908 vs. 0.794 and 0.874), and appears to yield more granular and easily visualized risk estimates.

1. Introduction

Intensive care in the United States consumes nearly 1% of the country’s GDP annually and accounts for 13% of all hospital costs, as of 2010. The human toll of ICU care in the U.S. also presents a large burden: nearly 6 million patients are admitted to the ICU annually, and mortality rates in ICUs are estimated to be about 12%, on average. In addition, large variations in the quality of ICU care are also observed between hospitals, and much of this variation remains even after adjusting for disparities in case mix. Taken together, these findings underscore the critical need for performance measurement and benchmarking of ICU outcomes, chiefly of mortality and length of stay. Such initiatives are driven by the application of accurate ICU risk adjustment models, including the Acute Physiology and Chronic Health Evaluation (APACHE), the Simplified Acute Physiology Score (SAPS), and the Mortality Probability Model (MPM). In addition, these models have also been used to adjust for case mix in observational studies, to compare treatment arms in RCTs, and to inform ICU decision-making, triage, and resource allocation.

However, existing ICU models exhibit a series of limitations. As the increasing adoption of EHRs in the United States and elsewhere promises to make a wider variety of data available for these models to ingest, recent work in this area has focused on merely adapting existing models, including, for example, SAPS and APACHE to the EHR. However, in doing so, such approaches centered around adaptation neglect the wealth of information available in the form of unstructured, free-text clinical narratives, as well as trends in longitudinal laboratory test results and vital signs, possibly limiting predictive accuracy. If not EHR-adapted, these models then require manual chart abstraction by trained staff; one study estimated this abstraction process to take 37 and 20 min per chart, on average, for APACHE and SAPS, respectively. The time and cost burden associated with chart abstraction limits these models’ utility for real-time ICU decision-making, and makes it a nontrivial task to compute these risk scores retrospectively; in one study, the cost to abstract the data required to compute APACHE scores for a 60,000-patient ICU cohort was estimated to approach $2 million.

In this work, we present steps towards “fully electronic” ICU risk adjustment based on generalized additive models (GAMs) utilizing features built from fused clinical data to predict in-hospital mortality. By fused clinical data, we denote data which combines both structured and unstructured data sources; here, the former are derived from longitudinal laboratory test results and vital sign measurements from ICU flowsheets, while the latter derive from the free text of clinical narratives. Fused data of this form have previously been used to predict code status and colorectal surgical complications, but have yet to be used to predict ICU mortality or to estimate risk in the clinical setting more generally, aside from one
instance where topics derived from latent Dirichlet allocation were used in conjunction with SAPS for this task.16 Indeed, the use of fused data in clinical predictive modeling more generally appears to be rare, with nearly all published models leveraging either only structured or unstructured data sources, but not both in a single model.17 Such an approach, especially when combined with a flexible and interpretable method such as GAMs, could yield a richer model giving more personalized risk estimates obtained by the interaction of features derived from both types of sources. For example, trends in a patient’s Glasgow Coma Score (GCS) during their first 24 hours in the ICU could be interacted with mentions of postoperative status in provider notes to personalize their risk estimate to an extent not possible if relying on changes in GCS alone.

2. Methods

The data used in this study were drawn from the Medical Information Mart for Intensive Care-III (MIMIC-III) ICU database, developed from EHR, telemetry, and other data routinely collected for patients admitted between 2001 and 2012 at Beth Israel Deaconess Medical Center in Boston, Massachusetts, a tertiary care and teaching hospital.18 For this study, we identified the first ICU stay lasting > 4 h for each patient and selected all data that were collected up to 24 hours following ICU admission representing the laboratory tests and vital sign measurements listed in Table 1. This is in line with other ICU models, including APACHE IV and SAPS II, which look up to 24 hours following ICU admission; however, these models use only the ‘worst’ value, i.e., either the highest or lowest value within this window, depending on the variable. We also selected all provider notes, including physician progress notes, nursing notes, postoperative notes, and radiology reports written within this window, and did not differentiate on the basis of the author. The outcome used was in-hospital mortality.

To model the relationships between these features and mortality, we relied on an approach using generalized additive models (GAMs).19–21 GAMs represent an extension of generalized linear models (GLMs), where the predictors are related to the outcome via smooth, and possibly nonlinear, functions. The parametric forms of these functions can be prespecified before fitting the GAM and then the parameters estimated from the data, or otherwise be taken to be nonparametric and having arbitrary shape, with the latter approach being more common. Classes of functions commonly used in GAMs include locally weighted regression (LOESS) smoothers, smoothing splines, and regression splines. Here, we use smoothing splines, which afford greater flexibility than regression splines or LOESS smoothers, but tend to be more computationally expensive.

The GAM was compared to a series of models run against the fused dataset as described above, specifically $L_1$-regularized (or ridge) logistic regression, linear support vector machines (SVMs), and gradient boosted trees implemented using xgboost.22 We also obtained performance estimates for a logistic regression models based on both the SAPS and Oxford Acute Severity of Illness Score (OASIS)23 systems, which can easily be computed from data elements in MIMIC-3. We used the area under the receiver operating characteristic curve (AUC) was used a metric to compare models. Estimates of and 95% confidence intervals for the AUC for each model were obtained by 10 repetitions of 10-fold cross-validation (CV). All models and related procedures were implemented in R (version 3.3.3) using the caret and gam packages.

The physiologic data, comprising laboratory tests and vital sign measurements, were collected sequentially at varying sampling rates, and so each test or vital sign were initially represented by a time series. For example, heart rate was recorded hourly, while laboratory tests were taken less regularly. For each time series, in order to capture gross temporal variation, we engineered a set of derived features based on summary statistics, namely the mean value, standard deviation, maximum and minimum, last minus first value recorded as well as the absolute value of this difference, and the slope of the linear trend fit to the data. The full list of data elements and derived features used in our experiments is given in Table 1. We fit a GAM model using only these features derived from structured data as a baseline for the fused-data model.

For all fused-data models, prior to each fold of CV, and owing to computational considerations, we also separately fit a LASSO-based classifier, as in [24], to the free text narratives in the training set in order to prune the set of unstructured term mentions that eventually served as input to GAM, together with the features derived from structured data elements. On the order of $10^6$ unique terms were present across all the notes in our corpus, and this step resulted in roughly 500 unique terms being selected over each CV fold.
Without such a step, our GAM would have had to scale to accommodate on the order of \( O(\left[10^2+100\right]^2) \) possible interactions, a model fit which would have proved infeasible to compute. Throughout this process, each document comprising all the notes associated with a patient was represented as a bag of words without normalization, meaning that ontology mapping, negation detection, or other similar techniques were not used. Finally, the per-document frequencies associated with each term were then transformed into term frequency-inverse document frequency (tf-idf) \( ^{25} \) features which served as input to the GAM. We also performed a sensitivity analysis using the sublinear term frequency, i.e., \( \log(1+tf) \), in place of the raw term frequency in order to adjust for note length, which we hypothesized could be associated with mortality.

Table 1. List of structured data sources and the types of derived features engineered from each source. All derived features are with respect to a window of maximum length 24 hours following ICU admission.

<table>
<thead>
<tr>
<th>Laboratory tests</th>
<th>Vital signs</th>
<th>Derived feature types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood urea nitrogen</td>
<td>Heart rate</td>
<td>Mean</td>
</tr>
<tr>
<td>Bilirubin</td>
<td>Respiratory rate</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>Creatinine</td>
<td>Temperature</td>
<td>Maximum</td>
</tr>
<tr>
<td>Lactate</td>
<td>Mean arterial pressure</td>
<td>Minimum</td>
</tr>
<tr>
<td>Glucose</td>
<td>( S_o_2 ) (oxygen saturation)</td>
<td>Last value minus first value ( \Delta ) feature name</td>
</tr>
<tr>
<td>Sodium</td>
<td>( FiO_2 )</td>
<td>Absolute value of difference between last and first values</td>
</tr>
<tr>
<td>Potassium</td>
<td>Glasgow Coma Score (GCS) – total</td>
<td>Slope of linear trend fit to data using least squares</td>
</tr>
<tr>
<td>Bicarbonate</td>
<td>GCS – eye response</td>
<td></td>
</tr>
<tr>
<td>Hematocrit</td>
<td>GCS – motor response</td>
<td></td>
</tr>
<tr>
<td>White blood cell count</td>
<td>GCS – verbal response</td>
<td></td>
</tr>
<tr>
<td>Platelet count</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arterial ( P_aCO_2 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As the data used in this study were de-identified, this study was deemed to be exempt from review by the Institutional Review Board of the Stanford University School of Medicine.

3. Results

The characteristics of the dataset are presented in Table 2. The data reflect a rich case mix, with at least six different types of ICUs represented, including coronary care and cardiac surgery recovery units, medical ICUs, surgical ICUs, and a combined trauma/surgical ICU.

Table 2. Characteristics of the dataset.

<table>
<thead>
<tr>
<th>Patients, total number</th>
<th>36,043</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths (%)</td>
<td>3,895 (10.8%)</td>
</tr>
<tr>
<td>Age, mean (IQR)</td>
<td>61.9 (51-76)</td>
</tr>
<tr>
<td>Of which male (%)</td>
<td>20,836 (57.8%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of ICU</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary care</td>
<td>5,255 (14.6%)</td>
</tr>
<tr>
<td>Cardiac surgery recovery unit</td>
<td>7,394 (20.5%)</td>
</tr>
<tr>
<td>Medical (including Neuro ICU)</td>
<td>12,549 (34.8%)</td>
</tr>
<tr>
<td>Surgical</td>
<td>5,963 (16.5%)</td>
</tr>
<tr>
<td>Trauma/Surgical</td>
<td>4,882 (13.6%)</td>
</tr>
</tbody>
</table>

The results of various models are presented in Table 3. The GAM outperformed a logistic regression model based on SAPS, with 10-fold cross-validated estimates of AUCs of 0.908 for the GAM versus 0.794 for logistic regression on SAPS, and 0.874 for logistic regression on OASIS. In addition, the performance of the GAM compares well to those of other models tested on the fused dataset; while an AUC of 0.910 was estimated for the gradient boosting machine (GBM) – the best-performing model in terms of raw AUC --
this difference was not statistically significant at the 0.05 confidence level, as the 95% CI for the estimate of the AUC obtained for the GBM failed to exclude 0.908. Furthermore, in a sensitivity analysis, using the a normalized tfidf metric incorporating a sublinear term frequency did not improve performance compared to the raw tfidf metric for each unstructured input, and so all performance estimates presented are based on the use of the latter metric. Finally, a GAM fit on fused data appears to yield statistically significantly better performance compared to a GAM fit on features derived from structured data alone.

Table 3. Model performance comparison.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression on SAPS score only</td>
<td>0.794 (0.790-0.798)</td>
</tr>
<tr>
<td>Logistic regression on OASIS score only</td>
<td>0.874 (0.864-0.881)</td>
</tr>
<tr>
<td>Logistic regression on fused dataset (FUSED)</td>
<td>0.857 (0.841-0.872)</td>
</tr>
<tr>
<td>Gradient boosting machine on FUSED</td>
<td>0.910 (0.901-0.920)</td>
</tr>
<tr>
<td>Support vector machine on FUSED</td>
<td>0.873 (0.855-0.893)</td>
</tr>
<tr>
<td>Generalized additive model on structured features</td>
<td>0.853 (0.840-0.866)</td>
</tr>
<tr>
<td><strong>Generalized additive model on FUSED</strong></td>
<td><strong>0.908 (0.898-0.917)</strong></td>
</tr>
</tbody>
</table>

Examples of univariate risk curves estimated by the GAM for single features are presented in Figure 1. Bivariate risk surfaces estimated for pairs of features, where both derive from unstructured data are presented in Figure 2, while those estimated for pairs where one feature derives from structured data and the other from unstructured data are presented in Figure 3. The univariate risk surfaces in Figure 1 appear to more accurately recapitulate the nonlinear nature of the relationships between these features and mortality. While the GAM does not provide summary estimates of the relative influence of each feature as a GLM would by estimating a single coefficient, some of the most influential features, as given by the p-value for the significance of the smooth term, were the features derived from GCS scores, particularly the mean GCS over 24 hours, the difference in GCS (ΔGCS) over this window, and the slope of the linear trend fit to these scores. Other influential features included patient age, and their mean sodium, potassium, and lactate levels, the smooths of which are shown in Figure 1. Some of the most influential unstructured features, again selected on the basis of their p-value, included “expired”, “CMO”, “midline shift”, “posturing”, and “BMT” (bone marrow transplant).

The bivariate plots presented in Figures 2 and 3 exhibit some interesting properties. First, these plots can be used to assess linear dependence between the features, as can be seen in the plots in Figure 2, as well as in Figure 3A. Second, a “phase transition” can be observed in Figure 3B, where the relationship between the windowed mean of the mean arterial pressure (MAP) and a term mention related to vasopressor medication, here “pressors”, begins to exhibit linear dependence for mean MAPs below 65 mm Hg. Similar patterns were observed for other related term mentions, including “levophed”, and “dopamine”, among others.

Third, in Figure 3C, the contour lines appear to recapitulate the shape of the univariate curve estimated for ΔGCS in Figure 1C, but rotated clockwise by 90 degrees, and we also observe that the risk estimates for ΔGCS are further stratified on the basis of postoperative status, as measured by mentions of “POD” (post-op day) in notes. Similarly, patients can be stratified both on the basis of their mean GCS during their first ICU day and also of their having experienced an overdose (Figure 3D), and the relationship between the two features also appears to exhibit linear dependence, in that post overdose status appears to confer a protective effect and vice versa. The highest-risk region lies in the bottom left of the plot, corresponding to other ICU patients with low mean GCSs who were not post overdose status. We cover some of these points in more depth in the Discussion section.

Figure 1. Bivariate risk curves for four features derived from physiologic measurements common in existing ICU risk models: serum sodium, potassium, and lactate, as well as the Glasgow Coma Score (GCS). However, for sodium and potassium levels, we have taken the 24-hour windowed mean, and for lactate and GCS, we have taken the windowed 24-hour difference. The grey regions denote 95% confidence intervals estimated by the GAM, and a rugplot lies at the bottom of each figure, giving the distribution of the feature among all patients in the dataset.
Figure 2. Examples of bivariate risk surfaces estimated for pairs of features derived from unstructured data. Risk estimates are represented by a red (low risk) to white (high) spectrum; the green lines denote contours joining areas of the plot having equal risk.
**Figure 3.** Bivariate risk surfaces for pairs of features, where one derives from unstructured data (x-axis) and the other from structured data sources (y-axis).

4. Discussion

Here, we contribute a risk adjustment model for ICU outcomes based on generalized additive modeling applied to fused clinical data, which comprises features derived from both structured and unstructured data sources. Our work highlights the utility of GAMs for predicting ICU mortality: not only did GAMs appear to significantly outperform existing ICU risk modeling methodologies – SAPS and OASIS – they also appeared to perform just as well as the current “gold standard” of predictive modeling – gradient boosting machines (GBMs) – applied to fused datasets. GAMs also appear capable of robustly estimating complex risk surfaces in order to produce more personalized risk estimates for ICU patients, and to a granularity not possible with other classifiers, including GBMs. In particular, our work demonstrates the utility of features derived from unstructured text, especially in conjunction with more traditional structured features via an interaction. However, further work remains to be done in order to be able to fully integrate features derived from text into ICU models, especially in conjunction with more traditional structured features via an interaction. However, further work remains to be done in order to be able to fully integrate features derived from text into ICU models, specifically with regard to normalization, which could possibly include mapping the terms onto an ontology and applying negation detection, among other methods, and also with regard to minimizing possible biases introduced by some of these terms, including, e.g., by “expired,” and to a lesser extent, by “DNR” (do not resuscitate) and “CMO” (comfort measures only). Nevertheless, it appears clear that in this setting that these features provide additional value beyond that yielded by features derived from structured data elements, which should motivate their use in such models.

Overall, the GAM approach appears to provide more easily interpretable estimates of risk for both the univariate and bivariate cases. The estimated risk curves agree more readily with clinical intuition, compared to those that would be obtained via other models: a GLM, for example, would estimate a straight line through each plot in Figure 1, which would fail to capture the true shape of the nonlinear relationship
between these features and mortality. In addition, these risk curves superficially resemble those used by the Rothman Index (RI), but are distinct in that the RI curves rely on polynomial regression and were estimated separately for each feature before fitting the model, while our approach uses more flexible smoothing splines and estimates all curves and surfaces together in a single model, potentially facilitating scalability and portability. Furthermore, and more generally, the RI also relies on the use of a customized data collection instrument separate from the main patient record, while our approach leverages only those data routinely collected during the process of care.

Examining the estimated risk curves for mean serum sodium and serum potassium over the up to 24 hour window following to ICU admission yielded by the GAM, we observe that they are U-shaped, with the minimum risk lying within common reference ranges for these tests, as would be expected. Similarly, the estimated risk curve for ΔGCS, or the difference between last and first Glasgow Coma Scores (GCSs) within this up to 24 hour window, again agrees with clinical intuition – that steeper declines in a patient’s level of consciousness are associated with a poorer prognosis. However, due to the lack of extreme-valued data, the risk estimates in these cases may not be reliable, and this is reflected by the width of the confidence intervals estimated in these ranges. For example, the relationship between extremely high or low sodium values and mortality, as suggested by the univariate plot in Figure 1A, would almost certainly not be borne out in reality; more data would be required to be able to obtain more accurate estimates of the true relationship.

Of particular interest are the bivariate risk surfaces for pairs of features estimated by the GAM in Figure 2. In order to gain some intuition, we first examine the risk surfaces estimated for pairs of terms taken from unstructured clinical narratives that we would almost certainly expect a priori to occur together in notes, and hence to exhibit interaction in a GAM. Here, these pairs are ‘CABG’ (coronary artery bypass graft) and ‘CSRU’ (cardiac surgery recovery unit), as well as ‘POD’ (post-operative day) and ‘dilaudid’ (an opioid analgesic commonly given after surgery). The plots of each risk surface are given in Figure 2. Again, whiter colors represent higher risk, while redder colors correspond to lower risk, and the green contour lines join points having equal risk.

The gradients of the contours in each plot in Figure 2 are diagonal, indicating a linear interaction between the terms in each pair, as would be expected. A diagonal gradient implies such an interaction as it shows that the risk contributions of an increase of a certain magnitude in the value of either feature are both approximately equal. In addition, the estimated risk decreases smoothly from the lower left to the upper right of each plot, which agrees with existing knowledge that patients admitted to ICUs postoperatively generally are at lower risk of in-hospital mortality. The significance of these interactions lies in that they can characterize the extent to which one feature acts as a proxy for the other in such models, allowing sources of redundancy as well as novelty to be identified – which may potentially prove significant when working with features derived from heterogeneous data sources.

Furthermore, in Figure 2, the bivariate risk surface estimated for “mannitol” (a medication used to reduce cerebral edema) and “midline shift” also implies a linear interaction between these features, though not over the whole range of each, as can be observed in the other two plots in Figure 2. Here, the risk increases superlinearly along the diagonal region of the plot, and suggests a synergistic relationship between the two features, as can be observed from the steepness of the risk surface in this region, as implied by the increased widths between the contours. This observation suggests that patients experiencing cerebral edema to an extent requiring osmotic therapies, such as mannitol, are indeed at much higher risk than would be suggested by the presence in a provider note of either “midline shift” or “mannitol” alone (mannitol has other uses, e.g., as a laxative).

In particular, bivariate risk surfaces can be estimated for pairs of features where one derives from structured data and the other from unstructured data, as in Figure 3. The risk surface estimated for “DNR” (do not resuscitate) and patient age is presented in Figure 3A. It is known that increasing patient age is associated with use of DNR orders, and this is borne out here, as the diagonal gradients again suggest the presence of an interaction. In addition, risk also increases as both the number of DNR mentions and age increase. Moreover, Figure 3B also illustrates an interesting property captured by these GAMs: below a windowed mean value of mean arterial pressure (MAP) of 65 mm Hg, a “phase transition” is observed. At this point,
the risk surface appears to abruptly shift from a regime where little to no interaction exists between these features, to one which exhibits robust linear interaction: the lower the value of mean MAP, the more frequently mentions of “pressor” appear, and the mortality risk increases concomitantly. A similar pattern with the same “phase transition” is also observed for other terms relating to vasopressors, including “levophed” and “dopamine”, among others. The existence of such a “phase transition” within these estimated risk surfaces is significant in that it delineates specific regions of the feature space where one feature can act as a proxy for the other and further informs the use of these features in risk models.

There are several limitations associated with our study. First, even though the data reflect a rich case mix and were collected over a 11-year-long period, they are representative of only one institution, and the relative portability of ICU risk models based on GAMs, compared to other classifiers, to other sites is unknown. Indeed, owing to their flexibility, and hence propensity to overfit, it is certainly possible that GAMs may not prove as portable as other classifiers. Second, we did not apply normalization in the preprocessing pipeline, and so did not perform negation detection or mapping terms onto an ontology, nor were the computed tfidf metrics normalized to account for note length, but our sensitivity analysis found that the latter modification did not change performance significantly. However, we felt that the development and validation of a such a pipeline was outside the scope of this study, which aimed to principally establish the feasibility of fully electronic ICU risk models based on fused data, and to demonstrate the value of these unstructured features in such models. Finally, the only comparators representing existing ICU risk modeling methodology available to us were the SAPS-I and OASIS systems, which may offer somewhat limited performance compared to the state-of-art in scoring systems, including APACHE IV, that rely on a broader range of features and so capture more information about a patient’s physiological state, but currently require costly manual chart abstraction. Ideally, a model based on the APACHE IV score would have been used as a comparator, but the estimated costs of data collection for this cohort – upwards of $1 million, based on other studies11,12 – precluded us from doing so.

Moreover, the full extent of the utility of using features derived from clinical narratives to predict ICU and other outcomes is currently unknown. Several examples exist in the literature where features derived from nursing notes have been used to predict ICU mortality,10,24 but to the best of our knowledge, no models currently deployed for this task rely on features derived from unstructured text data. While including these features in models could, as we have shown, potentially improve the predictive performance and facilitate interpretability of such models, their use also presents several unique challenges, given that documentation patterns may vary substantially between institutions. First, the mention of certain terms associated with documentation of patient death, including, e.g., “expired,” proves sufficient to predict the outcome with certainty, and as such, may suppress the contribution of other predictors and introduce at least one source of bias, owing to variability in documentation patterns.

To a lesser degree, this phenomenon also occurs with other terms such as “DNR” (do not resuscitate) and “CMO” (comfort measures only); indeed, it is known that DNRs are associated with mortality.37 Other similarly influential terms include “sepsis”, and those associated with neurological exams indicative of comatose status, e.g., “corneas” (as in “corneas absent”) and “posturing”. While there may not exist substantial variation in documentation patterns with regard to patient death, several studies have demonstrated institutional and regional variation in DNR ordering patterns,28,29 and such variation could present a potential source of bias -- one which could act in concert with variation in documentation patterns, although the full extent of the latter type of variation is presently unknown. Second, with the knowledge of the high influence of certain terms in these models, providers may then bias their documentation behaviors in such a way to “game” a model, e.g., by repeating terms associated with high risk while writing notes so as to inflate estimated mortality risks for their patients, thus improving the apparent risk-adjusted performance of their ICUs. However, these implementation challenges could be mitigated with new informatics methods built into the preprocessing pipelines that ingest the data for use by these models. Mitigation strategies could involve intelligently filtering potentially biased term mentions in unstructured narratives, as well as methods to characterize variation between providers’ lexical styles in notes, or to clamp the contributions of such a subset of features in these models.
5. Conclusion

In this paper, we demonstrated the utility of an approach based on generalized additive models for ICU mortality prediction on fused clinical data which combines features derived from both structured and unstructured data sources. The GAM approach offers an unified framework which outperforms an existing modeling paradigm based on SAPS, and in fact yields performance comparable to gradient boosting machines, and which allows for the estimation of complex — yet easily interpreted risk surfaces for pairs of features. In particular, our approach demonstrates the value added by the inclusion of features derived from unstructured narratives to further stratify mortality risk, particularly in concert with features derived from structured data sources and engineered so as to capture temporal variation. This approach may hold value for models estimated and deployed in other settings beyond the ICU to produce more personalized risk estimates for patients. However, there remain implementation challenges in utilizing unstructured data sources to their fullest extent in such models, but these could be mitigated with the development of adjunctive informatics methods.

Acknowledgements

BJM is supported in part by the National Library of Medicine (NLM) training grant T15 LM007033; JHC is supported in part by NIH Big Data 2 Knowledge Award Number K01ES026837 via the National Institute of Environmental Health Sciences. 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 or of Stanford Health Care.

References


Feasibility of Homomorphic Encryption for Sharing i2B2 Aggregate-Level Data in the Cloud

Jean Louis Raisaro, MS\textsuperscript{1}, Jeffrey G Klann, PhD\textsuperscript{2,3,4}, Kavishwar B Wagholikar, MD PhD\textsuperscript{2,3,4}, Hossein Estiri, PhD\textsuperscript{3,3,4}, Jean-Pierre Hubaux, Dr-Eng\textsuperscript{1} and Shawn N Murphy, MD, PhD\textsuperscript{2,3,4}

\textsuperscript{1}École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; \textsuperscript{2}Partners Healthcare, Boston, MA, USA; \textsuperscript{3}Harvard Medical School, Boston, MA, USA; \textsuperscript{4}Massachusetts General Hospital, Boston, MA, USA

Abstract

The biomedical community is lagging in the adoption of cloud computing for the management of medical data. The primary obstacles are concerns about privacy and security. In this paper, we explore the feasibility of using advanced privacy-enhancing technologies in order to enable the sharing of sensitive clinical data in a public cloud. Our goal is to facilitate sharing of clinical data in the cloud by minimizing the risk of unintended leakage of sensitive clinical information. In particular, we focus on homomorphic encryption, a specific type of encryption that offers the ability to run computation on the data while the data remains encrypted. This paper demonstrates that homomorphic encryption can be used efficiently to compute aggregating queries on the ciphertexts, along with providing end-to-end confidentiality of aggregate-level data from the i2b2 data model.

Introduction

With the increasing digitalization of medical information, the ability to share data across clinical sites is indispensable in realizing the full promise of personalized medicine. Yet, because of privacy and security concerns, sharing clinical data across healthcare organizations is currently extremely difficult or impossible to realize. The number of data breaches is on the rise and there is significant public pressure\textsuperscript{1-3} for clinical sites to not share data stored in their repositories to ensure that patients’ information is adequately protected. As a result, the privacy risks associated with medical data sharing represent a real barrier to medical advancement, if effective privacy-preserving technologies are not adopted.

Existing tools designed to enable collaborative medicine, such as Informatics for Integrating Biology and the Bedside (i2b2)\textsuperscript{4} and TranSMART\textsuperscript{5}, lack the necessary security and privacy guarantees to be applied in untrusted environments. Their current deployments pose significant challenges in terms of both security (access control, accountability, data traceability) and privacy (trust management, confidentiality and resilience against inference attacks) that need to be addressed in order to reassure clinical sites and foster clinical data-sharing. The evolution of the regulation towards further guarantees (e.g., HIPAA in USA and the new GDPR in EU) also reflects these urgent needs.

In the last few years, the IT security and privacy community has made substantial progress in the development of new sophisticated tools, also known as “privacy-enhancing technologies (PETs)”\textsuperscript{6}, that aim at responding to these needs. PETs provide strong security and privacy guarantees for the management of sensitive data\textsuperscript{6}. Yet, due to the lack of awareness and their inherent complexity, the biomedical community has largely left PETs unutilized. So far, only a few isolated PETs-based systems have been deployed in medical operational settings\textsuperscript{7-10}. Therefore, it is essential to bridge this cultural gap, in order to address the privacy and security concerns affecting the medical community.

This paper addresses this challenge by demonstrating the use of PETs in clinical environments. We explore the feasibility of using emerging technology for secure computation, specifically homomorphic encryption, for securely and privately sharing aggregate-level data from the most widespread data model for clinical research,
notably i2b2. Particularly, we introduce a new model that efficiently uses homomorphic encryption to enable clinical sites to share i2b2 aggregate-level data with multiple researchers through an untrusted cloud provider without having to worry about potential leakages of sensitive information to untrusted third parties.

i2b2 aggregate-level data consists of patient counts for unique combinations of location, ontology concept and time, e.g., the total number of patients at site S, that got “Type I Diabetes” observed in the first quarter of 2017. Although aggregate-level data is not directly identifying, sites willing to share the data by using a public cloud, are concerned about unintended leakages of sensitive information to untrusted third parties. For example, hospital-specific utilization patterns can reveal hospital operations and budgeting, and aggregates with small counts can reveal individual patients. Indeed, the leakage of this information can severely damage sites’ reputation and jeopardize the privacy of their patients. Our solution ensures end-to-end protection of data confidentiality against any third party not in possession of the decryption key. Only a legitimate and authorized investigator can see the end result of a query while the confidentiality of the original data is always preserved.

Informatics for Integrating Biology and the Bedside (i2b2)

i2b2 is an open source clinical platform for enabling secondary use of electronic health records (EHR) used at over 150 clinical sites and covering more than 250 million patients’ records only in the US. i2b2 is designed to enable investigators to perform queries on an enterprise data repository in order to find sets of patients that would be of interest for further clinical research studies. It consists of a simple and flexible relational data model based on a “star schema” and a set of server-side software modules, called “cells”, which are responsible for the business logic of the platform and are organized in a “hive”. The data model stores, in a narrow table called “observation fact” table, clinical observations (or “facts”) about patients such as diagnoses, medications, procedures, and demographics, along with a date, a patient identifier and an encounter identifier. Each observation is encoded by an ontology concept from a medical terminology, such as the International Classification of Disease (ICD) or the US National Drug Code (NDC). The use of extendable ontologies makes i2b2’s model highly adaptable to site-specific coding and easily deployable on top of existing EHR systems. Besides the observation fact table, there are four other “dimension” tables that further describe patients’ data and meta-data. Queries are built in a Web-based query tool by combining ontology codes, organized in a hierarchical tree-based structure, with logical ORs and ANDs operators. i2b2 is supported by an active community and has become one of the most popular open-source projects in health care research.

Homomorphic Encryption

Homomorphic encryption is a special form of public-key encryption that enables computations on encrypted values (see Figure 1). As with every public-key encryption scheme, it consists of two algorithms: an encryption algorithm and a decryption algorithm. Encryption and decryption can be seen as inverse operations. At one end, the encryption algorithm takes as input a plaintext message \( m \) and a public cryptographic key \( K \) and outputs a ciphertext \( c \). On the other end, the decryption algorithm takes as input the ciphertext \( c \) and a secret cryptographic key \( k \) and outputs the original message \( m \).

In general, the public key \( K \) is derived from a secret key \( k \) in such a way that the inverse operation is not possible. We can say that an encryption scheme is secure if the ciphertext does not reveal any information about the underlying message to anyone without the secret key \( k \) or, similarly, if anyone without the secret key \( k \) is unable to distinguish an encryption of a message \( m_0 \) from the encryption of another message \( m_1 \).
Figure 1. High-level representation of homomorphic encryption enabling computations on encrypted data. Different keys are used to encrypt and decrypt messages.

In contrast to standard public-key encryption that requires decryption in order to carry out any computation on the original messages, homomorphic encryption allows to perform computations directly on the encrypted messages. In fact, homomorphically-encrypted messages can be combined together, thus generating an encrypted result that, when decrypted, yields the result of a meaningful operation (like addition or multiplication) on the original messages. For example, by using the homomorphic encryption scheme ElGamal on elliptic curves\(^1\), also denoted as EC ElGamal, one can decrypt the sum of two ciphertexts and obtain the result of the sum of the two original messages.

Thanks to this property, homomorphic encryption schemes enable several use cases such as performing analytics in an untrusted environment (e.g., a public cloud) while ensuring the confidentiality of the data processed. Yet, it is important to note that homomorphic encryption introduces sometimes unpractical costs. For example, fully homomorphic encryption allows any computation to be carried out on the ciphertext but at the cost of unacceptable storage and computational overheads. Instead, partially homomorphic encryption has less flexibility than fully homomorphic encryption, as it allows only some specific computations to be carried out the ciphertext, but with reasonable storage and computation overheads. In this paper, we explore how to use additively (hence partially) homomorphic encryption, i.e., encryption that supports only additions on ciphertexts, to enable the sharing of aggregate-level data from the i2b2 data model.

**Methods**

Before introducing the proposed solution, we model the system and the threats that our solution aims at preventing.

**System and Threat Models**

We consider the data-sharing scenario depicted in Figure 2 where multiple clinical sites use i2b2 for internal purposes and want to share i2b2 aggregate-level data with multiple investigators by the means of a central public cloud. Several Clinical Data Research Networks (CDRNs) in the US National Patient-Centered Clinical Research Network (PCORNet)\(^1\) are concrete examples of such a data-sharing scenario. Actually, there is an increasing need to shift from decentralized data-sharing models to more centralized ones, because of insufficient resources (both human and technical) at sites to maintain an interoperable network.
In this setting, the standard system model usually consists of three main parties: (i) one or multiple clinical sites ($S_i$), organized in a network or federation, that aim at sharing their data through the cloud, (ii) one or multiple investigators willing to access this information, and (iii) a storage server (SS) hosted on the cloud and responsible for the central storage of sites’ data and for responding to investigators’ queries. Yet, in order to achieve strong security and privacy guarantees, we need to introduce an additional party on top of this model, namely a new server called proxy server ($PS$). The proxy server is necessary for helping the storage server during the privacy-preserving data-sharing protocol and ensuring trust decentralization. Trust decentralization is a crucial requirement for this kind of data-sharing scenarios as it allows to avoid single points of failure in the system. The storage and proxy servers can be represented by any pair of independent servers such as, for example, two servers at two distinct public cloud providers such as AWS Cloud or Google Cloud. It is crucial that these servers are independent and do not collude, otherwise the system could be easily compromised.

In this system model, we consider a semi-honest adversarial setting or threat model. This means that all the parties except the clinical sites, that are the legitimate owners of the data, are supposed to honestly follow the data-sharing protocol but might try to passively infer extra information from what they see during the protocol. For example, an insider or a hacker that has bypassed the cloud provider’s firewalls and tries to passively infer sensitive information from the data stored at SS, can be considered as a semi-honest adversary. Similarly, an investigator that legitimately queries the system with a series of carefully crafted requests in order to infer sensitive information about patients at the different sites, can also be considered as a semi-honest adversary. In contrast, an adversary that would try to cheat and depart from the protocol would be considered as malicious. Although, in most of the cases, adversaries are considered to be malicious, we note that the assumption of a semi-honest adversarial setting is reasonable in practice, especially in the medical context, as the different parties involved in the data-sharing process have either business/operational or reputational incentives to honestly follow the protocol and to not cheat. Hence, for the rest of the paper, we ignore the case of a malicious active adversary.
Data Model

In order to store i2b2 aggregate-level data, we designed a data model that is fully compatible with the i2b2 data model and consists of a single table with three aggregating attributes and one counting attribute (see Table 1). Each row of this table stores the total number of patients for whom the same observation (unique combination of location, ontology concept and time) has been registered in the i2b2 observation fact table of any of the sites. In particular, the location column stores the information about the site at which the observation has been collected. The ontology concept column stores the unique identifier (e.g., concept code or concept path from the hierarchical ontology tree) of the condition, medication, lab result, or procedure that has been observed. The time column stores the time unit at which the observation has been collected, e.g., a particular month, quarter, year. Finally, the count column stores the total number of patients with the observation specified by the ontology concept $X$ and collected at location $Y$ and time $Z$.

<table>
<thead>
<tr>
<th>Patients Totalnum</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK Location</td>
</tr>
<tr>
<td>PK Ontology_Concept</td>
</tr>
<tr>
<td>PK Time</td>
</tr>
<tr>
<td>Count</td>
</tr>
</tbody>
</table>

Table 1. Proposed data model storing i2b2 aggregate-level data.

This data model provides a high degree of flexibility as the three aggregating attributes can be defined at any spatial level, time scale and ontology concept of interest. Typical queries to this data model are of the form

\[
\text{SELECT } */\text{SUM(count)} \\
\text{FROM patients_totalnum} \\
\text{WHERE (ontology_concept=X_1 OR ontology_concept=X_2 OR ...)} \text{ AND} \\
(\text{location=Y_1 OR location=Y_2 OR ...}) \text{ AND (time=Z_1 OR time=Z_2 OR ...)} \\
\text{GROUP BY *;} \\
\]

where $X_i$, $Y_i$ and $Z_i$ denote any possible value in the corresponding column. A concrete example of such queries could be: “What is the total number of patients per site that had been diagnosed with HIV in 2015?”. Our privacy-preserving model enables queries of this type on homomorphically encrypted data.

Privacy-Preserving Data-Sharing Protocol

The proposed privacy-preserving data sharing protocol is described in Figure 3 and it is based on the additively-homomorphic properties of the EC ElGamal cryptosystem\(^\text{16}\).

**Initialization phase.** We begin with the initialization phase. During this phase, the storage and the proxy servers, both equipped with a processing unit named “crypto engine”, independently generate a pair of asymmetric cryptographic keys (public and secret keys) for the EC ElGamal cryptosystem and combine their public keys to obtain a public shared key (black key in Figure 3). The two servers store their secret keys safely and do not exchange them with any other party in the system.
Figure 3. High-level representation of the proposed privacy-preserving data-sharing protocol. Steps (A-C) represent the ETL phase whereas steps (1-5) represent the secure query processing.

**Extract, transformation and loading (ETL) phase.** In the ETL phase, the *shared* public key is used by each site to encrypt their i2b2 aggregate-level. More specifically, each site extracts plaintext aggregate-level data from their i2b2 local databases, transform it in order to fit the aggregate data model, encrypts the total numbers of patients and uploads the encrypted table to the storage server in the cloud. Encryption under the shared key ensures that as long as an adversary does not obtain the secret keys of both the *storage* and *proxy* servers, the confidentiality of the data is preserved. In this way, we avoid a single point of failure in the system as both cloud providers must be compromised simultaneously in order, for an adversary, to reconstruct the *shared* secret required to decrypt the encrypted data stored at the storage server.

**Querying phase.** Once the encrypted data is securely stored on the *storage* server, an investigator can query it by specifying the combination of aggregating attributes values for which he wants to receive the patients’ total numbers. As such, the investigator generates a temporary pair of asymmetric cryptographic keys (public and secret keys in red in Figure 3) and sends the query along with his public key to the *storage* server (step 1 in Figure 3). The temporary pair of keys is freshly generated for each new query. Once the query received, the *storage* server fetches from the database the records, encrypted under the *shared* public key, that satisfy the query criteria (step 2 in Figure 3) and uses the *crypto engine* to homomorphically aggregate them (step 3 in Figure 3).

Due to the additively-homomorphic properties of the EC ElGamal encryption scheme, records can be aggregated without the need of decryption, thus yielding an encrypted result. We note that a *shared* secret key corresponding to the *shared* public key does not exist in the system and decryption would not be possible unless the *storage* and *proxy* servers collude, which is highly unlikely based on the above-mentioned system model. Yet, in order to be decrypted by the investigator, the result encrypted under the *shared* public key must be re-encrypted under the investigator’s public key. Therefore, the *storage* server starts an interactive two-party *re-encryption* protocol with the *proxy* server in order to switch the encryption of the result (step 4 in Figure 3) so that it is decryptable only by the investigator issuing the query. As such, each server sequentially uses its secret key and the investigator’s public key to partially modify the encryption of the result.

At the end of the protocol, the *storage* server obtains the query result encrypted under the investigator’s key. Note that during the re-encryption protocol, none of servers can see the query result in the clear. Optionally and
depending on the trust level of the investigator, the storage server can then homomorphically add random Laplacian noise to obfuscate the query result and achieve differential privacy\textsuperscript{17,18} in order to mitigate the risk of re-identification of a patient in one of the sites’ databases.

Finally, the (potentially obfuscated) result encrypted under the investigator’s public key is sent back to the investigator who can decrypt it by using his secret key (step 5 in Figure 3) and visualize the result. We note that the use of differential privacy implies that during the system initialization phase, each investigator is assigned a differential privacy budget “ε” that is progressively reduced each time a new query is performed. When the budget is zero, the investigator cannot perform queries anymore.

Results

Prototype Implementation

We implemented the proposed solution as a client-server application in the Go language (version 1.8) by relying on the open-source ONet framework for decentralized and distributed protocols\textsuperscript{19}. For the cryptographic operations in the privacy-preserving data-sharing protocol, we relied on the Go’s native cryptographic libraries and the ONet advanced cryptographic library\textsuperscript{19} that includes an implementation of the EC ElGamal on the Ed25529 elliptic curve\textsuperscript{12} with 128-bit security.

The source code of our application is publicly available at https://github.com/JLRgithub/PDCi2b2. The client component enables the user to encrypt aggregate-level data extracted from the i2b2 data model and to specify the desired combination of aggregating attributes in order to query the encrypted data stored in a relational database system on the storage server. The server component enables to setup the storage and proxy servers, generate the cryptographic keys and execute the functionalities of the “crypto engine”, i.e., the homomorphic aggregation, the two party re-encryption protocol and the homomorphic obfuscation.

Performance Evaluation

We tested the performance of the proposed model in a real cloud-based environment. We set up the storage and proxy servers on two separated AWS instances both running 64-bit Linux Amazon-edition on a dual-core Intel Xenon E5645 processor with 2.40GHz frequency and 4GB RAM. Note that for an ideal deployment the two instances must be hosted by two different cloud providers in order to minimize the risk of collusion. We used an off-the-shelf laptop running Mac OS X on Intel Core i7 processor with 3.1GHz frequency and 16GB ram as the client machine.

We simulated three different sites and, for each of them, we encrypted patients’ total numbers extracted from the i2b2 demo dataset. We considered only concepts in the lowest level of the ontology tree-based hierarchy, i.e., the leaves, and we used quarters as time unit. Patient counts for ontology concepts at higher levels in the tree-based hierarchy can be obtained by recursively aggregating patient counts for the children concepts. Therefore, we used the concept path to uniquely identify each ontology concept in order to keep track of all the parental levels in the ontology tree. The resulting database had 10,767 rows per site for a total of 32,301 rows. We store the encrypted data in a PostgreSQL database on the storage cloud server.

We measured the time of each cryptographic operation used in the privacy-preserving data-sharing protocol by averaging over 100 random queries for different levels of aggregations and different combinations of ontology concepts, times and locations. We used a single thread of execution in all of our experiments, not taking advantage of the available parallelism, to obtain a lower bound on the performance of the proposed system. Table 2 shows the amortized times obtained in our experiments. We note, however, that these operations are naturally parallelizable hence performance can be significantly improved.
Table 2. Amortized times in milliseconds for the cryptographic operations run in the proposed privacy-preserving data-sharing protocol.

<table>
<thead>
<tr>
<th>Cryptographic Operation</th>
<th>Amortized Time per Record (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homomorphic aggregation</td>
<td>0.004</td>
</tr>
<tr>
<td>Proxy Re-Encryption</td>
<td>13.77</td>
</tr>
<tr>
<td>Decryption</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Homomorphic aggregation scales linearly with the number of records that satisfy the querying criteria, whereas the re-encryption and the decryption operations scale linearly with the number of query results. As an illustrative example, an aggregate query asking the total number of patients for whom any diabetes-related concepts have been observed in the first quarter of 2012 across all sites takes, on average, a total of 807 ms, where 214 ms is the time necessary to fetch the encrypted records satisfying the query (121 ontology concepts per site contain the keyword “diabetes” in their ontology path) from the database, 1.4 ms are used to homomorphically aggregate the records, 13 ms are used in the re-encryption protocol and 0.9 ms to decrypt the result. The remaining time, 576 ms, is the communication time between the client and the storage server necessary for sending the query and receiving the encrypted result, which was constant for all tested queries. Contrarily, a non-aggregate query that does not involve homomorphic aggregation, such as the total number of patients at site 1 for whom Diabetes Mellitus (ICD9:250) has been observed in the first quarter of 2012, takes only 590 ms in total (0.58 ms to fetch the record from the database, 13 ms for re-encryption, 0.9 ms for decryption and 576 ms for communication).

Discussion

In this paper, we demonstrated the feasibility of using homomorphic encryption in order to enable the sharing of i2b2 aggregate-level data through the cloud. We showed how this type of privacy-enhancing technology can be efficiently used to enable new use cases that are currently impossible due to technical or legal reasons, or need to rely on costly and time-consuming ethical and legal processes. Our solution introduces a new and simple model that minimizes the risk of unintended leakages of sensitive information by ensuring end-to-end confidentiality protection with respect to untrusted third parties. Results obtained in a basic testing environment are promising and show reasonable overhead especially from a computational point of view. The storage overhead is also practical as an encrypted database is only four times larger than its unencrypted version.

Yet, it is important to mention that a main limitation of the proposed model is availability— both the cloud servers (storage and proxy servers) need to be online in order for the system to be operational. Although this might slow down adoption in an operational setting, replication can be used to mitigate this problem. Actually, it would be sufficient to replicate the number of instances of the same servers at the two cloud providers in order to guarantee service continuity: if one instance is unavailable, another one can immediately replace it.

We emphasize on how the presence of two servers is necessary for the overall security of the system in order to achieve trust distribution and avoid a single point of failure in the system. If the cloud provider hosting the storage server was compromised, the data would still be secure because the attacker would need to also compromise the cloud provider hosting the proxy server in order to reconstruct the shared decryption key necessary to decrypt the data. Differently, if the security of the whole system was centralized on a single party (e.g., only on the storage cloud), it would be easier for the attacker to concentrate all his resources on a single target thus increasing the likelihood of jeopardizing the whole system. Intuitively, the larger the number of parties among which the trust is distributed (i.e., the number of proxy servers) the more secure is the system. Systems such as UnLynx\textsuperscript{20} or Sharemind\textsuperscript{21} rely on this idea. However, we believe that two independent servers can represent a practical and acceptable compromise between security and system usability.

Moreover, we note that the encryption scheme used (EC ElGamal) is only additively homomorphic, hence it enables only linear aggregations in the encrypted domain (e.g., summations, counts). For more complex types of
aggregation (e.g., correlations, linear regressions) a somewhat homomorphic encryption scheme, such as the FV encryption scheme\textsuperscript{22}, that also enables multiplications on encrypted data can be used instead of EC ElGamal at the expense of an increased storage overhead, without modifying the proposed model. Although we used homomorphic encryption, there are other privacy-enhancing technologies that could also be used in this context. Secure computation is well supported also by garbled circuits\textsuperscript{23}, as well as by linear secret sharing\textsuperscript{24}. Yet, these two technologies require an increased level of interaction between the different parties than homomorphic encryption, thus resulting in even less reliable models from an operational perspective.

In conclusion, the proposed solution based on homomorphic encryption has been shown to be efficient and secure in the semi-honest adversarial setting as sites’ sensitive data remains encrypted during the whole data-sharing process. Only the investigator sending the query is able to decrypt the end result. In this paper, we did not address the case of a malicious active adversary who tampers with the data and the protocol in order to infer sensitive information as it would be in contrast with the business incentives of a cloud provider. Yet, we note that our model can be easily extended to this case in order to protect also the integrity of the data and the computations carried out in the cloud. Techniques such as zero-knowledge proofs of correctness\textsuperscript{25}, can be added on top of our encryption scheme in order to ensure the integrity of the different steps of the data-sharing protocol and the identification of misbehaving parties at the expenses of an increased computational cost. These proofs can be securely stored on a distributed ledger, e.g., a blockchain, and be verified by any party in the system. We will consider this enhancement in future work.

Conclusion

Privacy and security concerns represent real obstacles for medical progress. Bridging the gap that still exists between the medical community and the IT security and privacy community is key for addressing this problem and making new and sophisticated privacy-enhancing technologies be core components of clinical operational tools in the near future. Although homomorphic encryption is still at an early point in its life cycle, there has been consistent, substantive, and rapid progress in making it practical from a performance standpoint. In this paper, we have shown that homomorphic encryption has enormous potential for enabling new use cases such as privacy-conscious sharing of i2b2 aggregate-level data in the cloud.

The proposed model is generalizable and is not tied to i2b2 aggregate-level data. In future work, we will extend its functionalities so that also individual-level data from the i2b2 data model could be shared in the cloud in a privacy-preserving way thus making a further step towards the realization of personalized medicine to its full potential.

References


19. DEDIS/EPFL. The Cothority Network Library - ONetNo Title.


Provider Adoption of Speech Recognition and its Impact on Satisfaction, Documentation Quality, Efficiency, and Cost in an Inpatient EHR

Kshitij Saxena, MD, MHSA,1 Robert Diamond,1 Reid F. Conant, MD, FACEP,2 Terri H. Mitchell, MSN, RN-BC,2 ir. Guido Gallopyn,2 Kristin E. Yakimow, MS, MBA2
1Health Quest, Lagrangeville, New York; 2Nuance® Communications, Inc., Burlington, Massachusetts

ABSTRACT
This study utilizes qualitative and quantitative methods to measure the adoption of speech recognition (SR) and its impact on provider satisfaction, documentation quality, efficiency, and cost when used for clinical documentation within the electronic health record (EHR). Qualitative surveys gauged providers’ expectations and experiences regarding documentation before and after SR implementation. A new methodology was developed to measure SR adoption as a proportion of total documentation volume. Quantitative data was collected from the EHR, medical transcription and SR solutions to measure SR adoption and cost savings. Study results revealed significant improvements in satisfaction, documentation quality, and efficiency among providers as a direct result of SR use. An improved provider experience correlated to an 81% reduction in monthly medical transcription costs, an increase from 20% to 77% in electronic clinical documentation adoption, and a 74% SR adoption rate.

Keywords: speech recognition, medical transcription, electronic health records, health information technology, electronic provider documentation, clinical documentation improvement, provider satisfaction

INTRODUCTION AND SIGNIFICANCE
The US health information technology landscape has experienced remarkable transformation. Nearly all hospitals have adopted a certified electronic health record (EHR), but many report barriers, according to new data from the Office of the National Coordinator for Health Information Technology (ONC). In May 2016, the ONC released data briefs on EHR adoption and interoperability based on the American Hospital Association’s annual survey. According to one brief, 96% of non-federal acute care hospitals have adopted a certified EHR, up from 71.9% in 2011, the year the federal EHR incentive program launched. Adoption of basic EHR functionality between 2008 and 2015 rose from 9.4% to 83.8%, with eight of ten small, rural, and critical access hospitals having also implemented a basic EHR.

Since the EHR incentive program began, the American Medical Informatics Association (AMIA) and the American College of Physicians (ACP) have released policy meeting position papers regarding the purposes of clinical documentation and challenges that have arisen from EHRs. According to the ACP, requirements for capturing structured data should be kept to a minimum, never take the place of meaningful narrative comments, and EHRs should be leveraged for improving documentation and care. AMIA meeting consensus revealed that in the transition to a more technology-enabled healthcare environment, the main purpose of documentation should be to support patient care and improve outcomes. Documentation for other purposes should be a byproduct of care delivery. Guiding principles for clinical data capture and documentation should support high quality information that is accurate, relevant, confidential, reliable, valid, complete, and secure while supporting downstream uses such as quality measurement, performance improvement, population health care delivery, policymaking, research, education, and reimbursement. Furthermore, central to a learning healthcare system, is the need for data capture and presentation methods that support clinicians’ cognitive needs and should align with, but not impede, the care team’s workflow and care delivery.

With clinician burnout and shortages on the rise, the stakes for improving clinical documentation, workflow, and EHR optimization are imperative. According to a discussion paper published by the National Academy of Medicine (NAM), challenges affecting clinical practice are new payment and delivery approaches, EHRs, patient portals, and publicly reported quality metrics, all of which change how care is provided, documented, and reimbursed. Furthermore, they express concern that more than half of physicians are experiencing substantial symptoms of burnout while nurses report similarly high prevalence of burnout and depression. Work process inefficiencies such as computerized order entry, administrative/clerical burden, and documentation are cited as key contributors to burnout.

As a solution to improve burnout, satisfaction, and other clinical documentation issues, the introduction and integration of speech recognition (SR) technology in EHR systems has made significant progress in the last decade. In 2004, an analysis of the implementation and impact of SR in healthcare concluded that, while SR already had a significant impact on the ability of healthcare providers to operate more cost effectively and provide better patient care, there was still much room for improvement in the way SR technology was adopted and implemented. A study, conducted in 2006, regarding SR acceptance by 186 physicians found experiences were more negative than expectations, particularly with respect to technical performance, medical record quality, and time spent on documentation. Lyons et al. (2015) conducted a replication survey of physicians’ expectations and experience with SR technology before and after implementation with emergency medicine physicians. The expectations survey was administered prior to SR training and the experience survey after six months of SR use. In this study, 82% of physicians were initially optimistic that using SR with the electronic record was a good idea. After using SR, 87% agreed that SR was a good idea. In addition, 72% of physicians expected that SR use would save time, while only 51% reported time savings. The increased acceptance of SR in this study was attributed to improvements in the technology and the electronic record. In 2014, a systematic review of SR technology in healthcare concluded that SR systems have substantial benefits and should be considered in light of costs and system selection, training requirements, transcription length, potential use of macros and templates, the presence of accentuated voices, and workflow patterns. Expectations, combined with the need for prolonged engagement, should be managed during implementation. Improved turnaround times of diagnostic procedure reports or similar tasks represent an important outcome as it impacts timely delivery of quality care. Finally, as technology advances, uptake of SR systems will increase by diverse health and support staff working within a range of healthcare settings. Through the identification of studies published prior to December 2014, Hodgson and Coiera (2016) conducted a literature review, focused on risks and benefits, to assess the impact of SR on clinical documentation. Results discovered that many studies compared SR to dictation and transcription in radiology and heterogeneity across studies was high. Dictation and document editing time increased using SR, while lower rates for clinically significant errors were reported for dictation, which was partially due to documentation being completed by skilled transcriptionists as an added safety check. Turnaround time consistently improved when using SR compared to
dictation across all studies, while SR accuracy also improved. Economic benefits were poorly reported. SR technology is steadily maturing and offers advantages for clinical documentation, however, evidence supporting the use of SR is weak and further investigation is required to assess the impact of SR on documentation error types, rates, and clinical outcomes. In 2015, a randomized control trial compared the impact of SR with self-typing based on measurement of documentation speed, volume, and user satisfaction. This study, based on 1455 medical documents, demonstrated that the availability of SR led to increased documentation speed, increased documentation amount, and higher physician satisfaction.

Ensuring efficient, accurate, timely electronic documentation capture and communication continues to be a key initiative by government bodies and influential national professional associations, thereby, an ongoing focus for healthcare organizations. Health Quest, the largest family of nonprofit hospitals and healthcare providers in Hudson Valley, New York, was no different from other organizations seeking to improve clinical documentation capture, while also seeking to increase adoption and provider satisfaction with clinical documentation and the EHR. Health Quest implemented Cerner® Millennium® in 2008 and achieved HIMSS Analytics® Electronic Medical Record Adoption Model®️ Stage 6 by 2011. While working to achieve Stage 7, Health Quest sought to improve efficient, timely, accurate, electronic provider documentation capture in addition to provider satisfaction as part of their strategic healthcare information technology priorities.

In 2013, providers were experiencing workflow challenges documenting notes, diagnoses, and treatments, as well as capturing the full clinical narrative in the EHR through traditional input modalities (i.e. handwritten scanned notes, back-end medical transcription, and electronic documentation through manual keyboard [KB] entry). Handwritten notes were scanned into the EHR in batches approximately once every four to six hours, creating delays in the documentation process, making information difficult to find and causing legibility concerns. Back-end, technology-enabled medical transcription was another input modality. Providers dictated patient notes through a microphone into medical transcription speech recognition technology, which captured the voice and transcribed it into a complete patient note through both SR technology and medical transcriptionists’ review. Although faster than handwritten scanned notes, this modality required approximately 90 minutes of turnaround time for notes to be available in the EHR (variation exists by note type, policy and procedure, contract requirements and can range from 24-48 hours depending on note type). Through the final input modality, providers used electronic documentation, typing their notes directly into the EHR with point-and-click functionality. This solution included the option to use documentation templates and macros based on the provider’s specialty.

Through the integration of SR technology, Health Quest leadership sought to improve provider adoption of and satisfaction with the EHR at each of their hospitals, including the 365-bed flagship hospital, Vassar Brothers Medical Center (VBMC) in Poughkeepsie, New York, in addition to overcoming estimated annual transcription costs exceeding $1 million. Health Quest’s primary goals were to drive more accurate, comprehensive electronic provider documentation through front-end speech recognition, improve provider satisfaction with the EHR and clinical documentation process, and increase adoption of the EHR.

In June 2014, Health Quest implemented a pilot of Cerner® Dynamic Documentation™ and Nuance® Dragon® Medical speech recognition technology [DD+SR] with five hospitalists at each hospital, including VBMC, to collect feedback and evaluate their implementation strategy. With SR, providers could speak their notes directly into the EHR in real-time, with greater detail, capturing both structured information and the unstructured clinical narrative. The expectation was that legibility issues, delayed access to clinical information for patient care, and the downstream impact of clinical documentation issues and delays (i.e., availability to nursing, case management, quality, coding, billing, and others) would be greatly reduced or eliminated.

Implementation of DD+SR consisted of technology components that enhanced electronic provider documentation, including the ability to dictate directly into the EHR using specialty-based document “templates” to create comprehensive notes. Providers could now click within a chosen field (e.g. history of present illness, past medical history, physical examination, or operative notes) and begin to dictate that section of the note. Additionally, the ability to speak-enable the EHR-based Smart Templates and auto-texts further enhanced documentation by allowing providers the option to retrieve patient exams, lab results, reports, and other EHR data, to incorporate them into the patient note along with clinical narrative. Finally, by deploying DD and SR together, providers could leverage the EHR “tagging” function, whereby sections of documentation could be tagged for simplified inclusion in reports.

Based upon pilot users’ feedback, the implementation team elected to use standardized, specialty-specific auto-text templates and Smart Templates, based on documentation best practices for the continued deployment of DD+SR. The pilot was expanded to all hospitalists in July 2014, intensivists in October 2014, and DD+SR was deployed to all VMBC provider specialties during the final phase of implementation beginning March 30, 2015, which resulted in more than 150 providers trained during that timeframe.

Health Quest partnered with Nuance to conduct a research study during the final phase of DD+SR implementation, at VBMC, with three key objectives in mind. First was to support requests by ONC, AMIA,ACP, NAM, and other organizations seeking to expand upon research and best practices regarding healthcare information technology, clinical documentation improvement, efficiency, workflow, and clinician satisfaction. Second was to expand upon requests for future research on SR noted in the review of literature. The third objective was to validate and share best practices with other healthcare organizations seeking to drive more accurate, comprehensive electronic provider documentation through front-end speech recognition, improve provider satisfaction with the EHR and clinical documentation processes, while promoting EHR adoption and decreasing documentation costs.

The 31-month study covered four periods: pre-implementation (before: January 6, 2014 – May 31, 2014), pilot (June 1, 2014 – March 29, 2015), implementation (transition: March 30, 2015 – May 1, 2015), and post-implementation (after: May 2, 2015 – August 1, 2016).

**METHODS**

A two-pronged study consisting of qualitative and quantitative segments was conducted. The qualitative segment used surveys to collect information on providers’ experiences and expectations regarding clinical documentation before and after deployment of SR to measure provider satisfaction, documentation quality, and efficiency. The quantitative segment collected data from the EHR, transcription, and SR solutions to measure note volume evolution per input modality. Additionally, the quantitative segment measured adoption and cost savings. All quantitative measurements were via retrospective, longitudinal, observational study.
Qualitative Survey Method, Instrument, Participants, and Procedure

An electronic survey questionnaire was developed and deployed to measure expectations and experience regarding clinical documentation prior to SR implementation and to evaluate the post-implementation documentation experience including SR. The provider survey instrument was a pair of related and/or overlapping questionnaires that focused on satisfaction, documentation quality, and efficiency (Table 1). The surveys were delivered in two stages. During the pre-implementation period, the survey measured providers’ current experience with clinical documentation and expectations of SR. During the post-implementation period, providers were surveyed regarding their experience with the use of SR to facilitate the clinical documentation process. All VBMC survey participants were voluntary and consisted of physicians, physician assistants, and nurse practitioners who were trained on and used SR after training. Respondents included a mix of employed, community, and contracted providers from various specialties. Providers were given the opportunity to complete an electronic pre-implementation questionnaire when they attended training. Post-implementation surveys were administered electronically at a minimum of nine months post-implementation through a maximum of 17 months post-implementation.

<table>
<thead>
<tr>
<th>Survey Criteria</th>
<th>Questions</th>
</tr>
</thead>
</table>
| Provider Satisfaction| How likely are you to recommend documentation tools (pre) currently used in the EHR, (post) SR in combination with EHR to document the patient encounter to a colleague – on a scale from 0 to 10?  
(pre) Is it a good idea (post) Was it a good idea to introduce SR via Dragon Medical for documenting in the medical record: completely agree/somewhat agree/neutral/somewhat disagree/completely disagree? |
| Documentation Quality| After the introduction of SR, the quality and completeness of my documentation required for optimal care (pre) will (post) has: improved a lot/improved somewhat/remained the same/declined somewhat/declined a lot? 
After the introduction of SR, the time and effort I spend answering questions or clarifications from clinical documentation improvement (CDI), nursing, medical records, quality & coding staff (pre) will (post) has: decreased a lot/decreased somewhat/remained the same/increased somewhat/increased a lot? |
| Provider Efficiency | SR (pre) will (post) has improved/optimized workflow related to clinical documentation: completely agree/somewhat agree/neutral/somewhat disagree/completely disagree?  
After the introduction of SR, the time spent documenting the patient encounter (pre) will (post) has: decreased a lot/decreased somewhat/remained the same/increased somewhat/increased a lot? |

Table 1: Survey Questions

To study provider satisfaction, researchers used Net Promoter Score methodology. Respondents were asked, on a scale from 0 to 10, how likely they were to recommend current documentation tools versus SR in combination with the EHR for documenting in the medical record. The NPS methodology groups respondents as follows: Promoters (score 9-10) are loyal enthusiasts who will refer others. Passives (score 7-8) are satisfied, but unenthusiastic. Detractors (score 0-6) are unhappy and can impede use or growth through negative word of mouth. NPS represents number of Promoters minus number of Detractors, divided by total number of respondents, times 100. A positive NPS indicates more Promoters than Detractors; an NPS of +50 is considered excellent. Confidence intervals on NPS and significance of the difference of the pre- and post-implementation NPS were calculated using an established methodology.

To further evaluate satisfaction, provider skepticism regarding SR was assessed pre-implementation by asking whether it is a “good idea to introduce SR for documenting in the medical record.” Respondents could express their level of agreement or disagreement as follows: completely disagree, somewhat disagree, neutral, somewhat agree, completely agree. The five levels were translated with “somewhat agree” and “completely agree” mapping to 1, and the others to 0 to calculate binomial proportions. Post-implementation, to assess satisfaction regarding SR, the survey asked if it was a “good idea to introduce SR for documenting in the medical record.” The same level of agreement or disagreement options and calculation method was used.

To evaluate the impact on documentation quality required for optimal care with SR, the survey asked about expected improvement in documentation quality and completeness in the pre-implementation survey; and experienced improvement in documentation quality and completeness in the post-implementation survey. Respondents could express level of improvement with SR as follows: decline a lot, decline somewhat, remain the same, improve somewhat, improve a lot. The five levels were translated with “improve somewhat” and “improve a lot” mapping to 1, and the others to 0 to calculate binomial proportions.

To further evaluate documentation quality and efficiency, the survey assessed the time providers spent answering questions and clarifications from clinical documentation improvement (CDI), nursing, medical records, quality, and coding staff, with the use of SR. The pre-implementation question asked about expectations and the post-implementation question asked about the actual impact SR had on time spent answering questions and clarifications. Respondents could express level of time as follows: decrease a lot, decrease somewhat, remain the same, increase somewhat, increase a lot. The five levels were translated with “decrease somewhat” and “decrease a lot” mapping to 1, and the others to 0 to calculate binomial proportions.

To assess expected versus actual impact on provider efficiency with SR, respondents were asked to express their level of agreement or disagreement with whether SR improves clinical documentation workflow as follows: completely disagree, somewhat disagree, neutral, somewhat agree, completely agree. The five levels were translated with “somewhat agree” and “completely agree” mapping to 1, and the others to 0 to calculate binomial proportions.

To further evaluate impact on provider efficiency with SR, the survey asked about expected increase or decrease in time documenting the patient encounter with SR in the pre-implementation survey; and asked about time experienced documenting the patient encounter with SR in the post-implementation survey. Respondents could express time spent documenting the patient encounter with SR as follows: decrease a lot, decrease somewhat, remain the same, increase somewhat, increase a lot. The five levels were translated with “decrease somewhat” and “decrease a lot” mapping to 1, and the others to 0 to calculate binomial proportions.

To calculate 95% confidence intervals on the binomial proportions, the Wilson score interval method was used as recommended by Agresti and Coull.
Some respondents completed both surveys, while others completed either the pre- or post-implementation survey. The methodology proposed by Bland and Butland\textsuperscript{19} was used to compare the overlapping pre-implementation and post-implementation survey groups, to calculate the difference in proportions between the two groups, to calculate the confidence interval on the difference, and to perform a hypothesis test of equality of binomial samples with overlap. The results for the subset of respondents who participated in both surveys are presented as well; McNemar's Chi-squared test\textsuperscript{17} was used to determine if experience is consistent with the expectation for this group.

**Quantitative Methodology, Data Collection, and Data Processing**

The quantitative segment is a retrospective, longitudinal, observational 31-month study, from January 6, 2014 through August 1, 2016. During this study, VBMC transitioned from handwritten scanned notes and medical transcription to electronic provider documentation within the EHR.

A goal of this study was to measure how EHR note volume evolved over time for various input methods of clinical documentation: scanned handwritten notes, medical transcription, EHR data entry with keyboard and mouse and EHR data entry complemented with SR use. The evolution and degree of adoption of the EHR with the SR input modality and its effect on the volumes of other input modalities was of particular interest in this study.

A major challenge encountered during this study was the EHR did not capture if a provider note was created with or without SR. To overcome this measurement challenge, a new methodology was developed to recognize whether a note was created with SR or not at the scale of the large note volume (more than 1 million) observed during the study. The methodology is based on the detection of co-occurrence of SR sessions and EHR note change events for the same provider simultaneously active in both the EHR and SR systems. SR sent utilization packages to its database, in ten-minute intervals, when the microphone was on and consisted of speech duration in seconds, number of words recognized, and number of commands spoken. The EHR logged changes to a medical record in its database, which consisted of the following events: when a user opened the GUI wizard to create a new note, when a note was first saved, when a note was edited, when a note was saved again (could be repeated), and when a note was signed.

**Data Collection:** Custom SQL scripts were developed to retrieve user information; report information; the utilization packages from SR, EHR, and medical transcription systems; and note change-event log information from the EHR for the study period. These collections resulted in 18.28 gigabytes of raw data with 143,352,168 observations.

**Data Processing and Interlinking:** Users in the EHR and SR datasets and subjects in the survey datasets were interlinked using first and last names based upon the following methodology: A unique identification (ID) was established for 2,307 subjects, interlinking user account information in the EHR, medical transcription (459 users), SR (429 users), and the survey groups. Consecutive SR utilization packages by the same user were aggregated into a “session”, including quantification of the total number of words, total seconds of speech, as well as a session begin time and end time corresponding to when the microphone was turned on and off. Sessions without dictated speech were removed, resulting in 320,390 sessions in the study period. EHR report and action data were cleaned and merged in a reports dataset and a report event dataset.

EHR provider notes and SR sessions were interlinked with an algorithm developed for this study based on the following scenario: A provider logged into the EHR, created a new note, activated the microphone, dictated with SR in the EHR, saved and signed the note, and then turned off the microphone. This scenario would lead to the interlinking of the creation of a note in the EHR with the concurrent use of SR. Alternatively, sometimes, the provider created and subsequently saved and/or signed a note in the EHR without the concurrent use of SR. In this case, there would not be interlinking of the use of SR with the creation of that note. There were many possible variations of the exact user actions and sequence of the actions, but concurrent use of SR during the interval between creation and save/sign action of a provider note, within the same EHR session, led to interlinking indicating the use of SR as a part of the effort to create that provider note.

The SR-related user actions were captured in the SR dataset as a session, defined as the time interval between when the microphone was turned on and when it was turned off, with a start time, \( t_{\text{begin}} \), an end time, \( t_{\text{end}} \), the number of words dictated, all assigned to the respective SR user, during the time interval \( [t_{\text{begin}}, t_{\text{end}}] \). Likewise, the EHR-related user actions were captured in the EHR dataset as a sequence of “create,” “perform,” “modify,” “verify,” and “sign” note events with corresponding time-stamps. The “note create” event during a SR session was by itself not an indicator that the note was created with SR, as notes could be created and discarded without saving. Only EHR notes that were saved during a valid SR session were determined to have been created in conjunction with SR. As such, the interlinking methodology used the “perform” and “modify” events that indicate when a note was saved in the EHR.

Formally, researchers interlinked the SR session \( S \) and the EHR note \( N \), and classified \( N \) as created in conjunction with SR if the following conditions were true:

\[
\text{DragonTouch}(N_i) \equiv \exists S \in \text{SRSessions: Interlink } (N_i, S) \\
\text{InterLink } (N_i, S) \equiv \exists t_{\text{save}} \in \text{Save}(N_i); t_{\text{save}} < t_{\text{end}} & \text{ User}(N_i) = \text{ User}(S) \\
\text{Save}(N_i) \equiv \{ t_{\text{save}_1}, t_{\text{save}_2}, \ldots, t_{\text{save}_n} \}
\]

An efficient stack decoder algorithm in time complexity \( O(2N_{\text{sessions}} + M_{\text{events}}) \) to interlink EHR notes and SR sessions was developed based on the R package rstackdeque. EHR report data, SR data, medical transcription data, and interlinking information were merged into one large dataset. The dataset was de-identified; personal health information (PHI) elements and all persons’ names and related IDs were obfuscated.

The R statistical software environment and RStudio integrated development environment, in conjunction with the dplyr, tidyr, readr, RecordLinkage, rstackdeque, digest, ggplot, tables, knitr packages, were used for data engineering, exploratory analysis, statistical analysis, and report generation.
RESULTS

Qualitative Survey Findings

During DD and SR training, 121 providers completed the electronic pre-implementation survey. Post-implementation, 108 surveys were completed, resulting in an average of 15 months of SR use for those completing post-implementation surveys. 53 providers completed both pre- and post-implementation surveys. For the 108 post-implementation surveys, provider participation consisted of the following: 64% physicians, 18% nurse practitioners, and 18% physician assistants.

Qualitative results, summarized in Table 2, display findings for both surveys, including the subset of 53 providers who participated in both pre- and post-implementation surveys:

<table>
<thead>
<tr>
<th>Introduction of SR Survey Criteria</th>
<th>PRE</th>
<th>POST</th>
<th>Difference</th>
<th>Overlap Hypothesis Test (p-value)</th>
<th>PRE</th>
<th>POST</th>
<th>McNemar's Chi² test (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good idea to introduce SR</td>
<td>95% [64, 80]</td>
<td>22% [13, 31]</td>
<td>64% [64, 87]</td>
<td>77% [90, 100]</td>
<td>98% [62, 85]</td>
<td>0.0023</td>
<td></td>
</tr>
<tr>
<td>Documentation quality and completeness</td>
<td>69% [60, 76]</td>
<td>11.5% [19, 21]</td>
<td>64% [58, 82]</td>
<td>72% [62, 85]</td>
<td>75% [62, 85]</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>Decreased time spent answering questions &amp; clarifications</td>
<td>49% [40, 58]</td>
<td>11.6% [0, 23]</td>
<td>60% [31, 67]</td>
<td>55% [31, 67]</td>
<td>70% [56, 80]</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>Improved/optimized clinical documentation workflow</td>
<td>57% [48, 65]</td>
<td>26.4% [16, 36]</td>
<td>60% [47, 72]</td>
<td>60% [47, 72]</td>
<td>83% [71, 91]</td>
<td>0.00023</td>
<td></td>
</tr>
<tr>
<td>Decreased time spent documenting the patient encounter</td>
<td>46% [38, 55]</td>
<td>9.8% [-2.2, 22]</td>
<td>57% [48, 66]</td>
<td>55% [41, 67]</td>
<td>58% [45, 71]</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Qualitative Survey Results

Responses on likelihood to recommend documentation tools currently used in the EHR (pre-implementation) and SR in combination with the EHR (post-implementation) to document the patient encounter, grouped using Net Promoter Score methodology, are represented below (Table 3):

<table>
<thead>
<tr>
<th>Survey</th>
<th>N</th>
<th>Detractors</th>
<th>Passives</th>
<th>Promoters</th>
<th>NPS</th>
<th>SE</th>
<th>LCL</th>
<th>UCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Implementation</td>
<td>121</td>
<td>89</td>
<td>20</td>
<td>12</td>
<td>-63.6</td>
<td>5.96</td>
<td>75.3</td>
<td>52.0</td>
</tr>
<tr>
<td>Post-Implementation</td>
<td>108</td>
<td>16</td>
<td>38</td>
<td>54</td>
<td>35.2</td>
<td>6.97</td>
<td>21.5</td>
<td>48.8</td>
</tr>
<tr>
<td>Difference</td>
<td>n/a</td>
<td>-73</td>
<td>18</td>
<td>42</td>
<td>98.8</td>
<td>9.17</td>
<td>80.9</td>
<td>116.8</td>
</tr>
</tbody>
</table>

Table 3: Net Promoter Score

Provider Satisfaction: Survey results revealed that before the implementation of SR, documentation tools had an NPS of -64. The strongly negative NPS, with a sufficiently narrow confidence interval (CI) of [-75, -52], confirms provider dissatisfaction with prior documentation tools. After implementation, however, SR in combination with the EHR, had an NPS of +35, representing a 99-point positive shift in NPS post-implementation. This shift in NPS, with CI of [81, 117], indicates a dramatic increase in provider satisfaction that can be attributed to SR. The difference is statistically very significant as indicated by the very small p-value of the hypothesis test and the confidence interval on the difference. The result of all participants is consistent with the result in the subset of providers who participated in both pre- and post-implementation surveys. That is, the estimated NPS for the 53-person subset falls in the confidence interval estimated for all participants.

Expanding on provider satisfaction with the use of SR for electronic clinical documentation, the survey asked providers whether they believed it was a good idea to introduce SR for documenting in the medical record. Before implementation, expectations were high; 73% of providers, CI of [64, 80%], thought it was a good idea to introduce SR. These high expectations were exceeded as 95% of respondents, CI of [90, 98%], agreed in the post-implementation survey that implementing SR was a good idea. The p-value of the hypothesis test shows the increase of 22 points is statistically significant. The result of all participants is, again, consistent with the result in the subset of providers who participated in both surveys, as the results of the paired group falls in the confidence interval estimated for all participants.

Documentation Quality: Prior to SR implementation, 69% (CI [60, 76 %]) of providers expected to see an improvement in documentation quality and completeness required for optimal care with the introduction of SR. However, post-implementation, 81% (CI [73, 88 %]) of providers reported an actual improvement in quality and completeness of documentation required for optimal care while using SR. The difference between expectation and experience is 11.5 points, (CI [1.9, 21]), which is statistically significant for all participants. The subset of providers participating in both pre- and post-implementation surveys had expectations and experiences regarding documentation quality and completeness that were consistent with the results of all participants. Point estimates fall in confidence intervals for all participants. Yet, there is no significant difference in expected versus observed documentation quality and completeness in the subset of providers completing both surveys.
To expand upon documentation quality and efficiency, findings revealed 60% (CI [51.69] %) of providers experienced a decrease in time spent answering questions and clarifications from CDI, nursing, medical records, quality, and coding staff, whereas only approximately half (49% CI [40.58]) %) of providers expected to see a decrease. The difference between expectation and experience is borderline statistically significant. In the subset of 53 providers completing both surveys, the difference is estimated to be 15%, but is not statistically significant.

**Efficiency:** While 57% (CI [48.65]) %) of providers expected SR to improve/optimize workflow, 84% (CI [76.90]) %) of providers experienced improved/optimized workflow after SR was implemented. Experience exceeded expectations by 26 (CI [16, 36]) points, which is statistically very significant. The result of all participants is consistent with the result in the subset of providers who completed both surveys.

Expanding on efficiency, 57% (CI [48.66]) %) of providers reported experiencing a decrease in time spent documenting the patient encounter when using SR, while only 46% (CI [38, 55]) %) expected SR to decrease documentation time. While the difference of 9.8% is not statistically significant, the result of all participants is consistent with the subset who completed both surveys. Therefore, it can be concluded that most providers observed decreased time spent documenting the patient encounter with SR which was consistent with their expectations.

**Quantitative Results**

This study analyzed six different note types (History and Physical [H&P], Consults, Progress Notes, Anesthesia Notes, Surgery & Procedure Notes, and Discharge Notes) and five input modalities (Dynamic Documentation [DD+KB], DD+SR, Medical Transcription [MedTrans], PowerNote without SR, and Scanned Handwritten Notes [ScanDoc]) during the study period. Table 4 provides a breakdown of more than 1 million notes collected in the dataset, after processing and de-identification, as explained in the Methods section. Anesthesiologists at VBMC, using PowerNote, chose to continue using it in the EHR and did not migrate to DD+SR at the time of implementation.

<table>
<thead>
<tr>
<th>Note Type</th>
<th>All</th>
<th>DD+KB</th>
<th>DD+SR</th>
<th>Input Modality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anesthesia Note</td>
<td>143262</td>
<td>4</td>
<td>0</td>
<td>9874</td>
</tr>
<tr>
<td>Consultation Note</td>
<td>117736</td>
<td>14210</td>
<td>53053</td>
<td>39633</td>
</tr>
<tr>
<td>Discharge Summary</td>
<td>67201</td>
<td>10624</td>
<td>27056</td>
<td>22713</td>
</tr>
<tr>
<td>History and Physical</td>
<td>126039</td>
<td>10259</td>
<td>29351</td>
<td>14662</td>
</tr>
<tr>
<td>Progress Note</td>
<td>453711</td>
<td>92813</td>
<td>230484</td>
<td>0</td>
</tr>
<tr>
<td>Surgery &amp; Procedure</td>
<td>162510</td>
<td>4449</td>
<td>14445</td>
<td>19236</td>
</tr>
<tr>
<td>All</td>
<td>1070459</td>
<td>132359</td>
<td>354389</td>
<td>96244</td>
</tr>
</tbody>
</table>

**Table 4: Number of Notes per Note Type and Input Modality in the Dataset for the Entire Study Period**

**Note Volume Evolution:** Table 5 provides a breakdown of the mean note volume, measured in notes per week, across all input modalities by note type for the four study periods. Overall, there is a strong increase of 62.5% in note volume from 5873 (mean) +/- 276 (standard error) notes per week to 9546 +/-133 notes per week, with the increase being mostly pronounced in the post-implementation (after) period. H&P, Surgery and Procedure Notes, and Anesthesia Notes show only small variations across the study periods. Discharge Summaries increased 22% from 448.9 +/- 7.9 to 547.5 +/- 10.6 notes per week. Consultation Notes increased 47.6% from 688.6 +/- 59.9 to 1016.3 +/- 15.5 notes per week, and remarkably, Progress Notes increased 244% from 1406.6 +/- 112.9 to 4842.1 +/- 69.8 notes per week. This strong increase in Progress Note volume was not expected by the research team, considering the variation observed in H&P and Discharge Summaries that give an indication of the number of admissions and discharges per week. After consulting with the Health Information Management department, researchers found that Progress Notes were predominately handwritten and scanned into the electronic health record prior to implementation of DD+SR, but were correctly captured in the dataset. It is possible some of this variation may have been due to multiple provider notes on a single handwritten page being scanned in as a single document, whereas post-intervention these documents would be created as separate and multiple documents. At this time, however, the reasons for this unexpected outcome remains fully unexplained, a finding that future researchers may be interested in exploring further.

<table>
<thead>
<tr>
<th>Note Type</th>
<th>Before Mean ± Std Err</th>
<th>Pilot Mean ± Std Err</th>
<th>Transition Mean ± Std Err</th>
<th>After Mean ± Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anesthesia Note</td>
<td>1139.8 ± 59.9</td>
<td>1088.3 ± 37.6</td>
<td>919.8 ± 103.7</td>
<td>996.2 ± 18.0</td>
</tr>
<tr>
<td>Consultation Note</td>
<td>688.6 ± 15.6</td>
<td>722.9 ± 21.1</td>
<td>734.0 ± 64.9</td>
<td>1016.3 ± 15.5</td>
</tr>
<tr>
<td>Discharge Summary</td>
<td>448.9 ± 7.9</td>
<td>434.2 ± 13.0</td>
<td>420.0 ± 38.0</td>
<td>547.5 ± 10.6</td>
</tr>
<tr>
<td>History and Physical</td>
<td>966.3 ± 37.0</td>
<td>912.4 ± 26.7</td>
<td>780.2 ± 78.4</td>
<td>921.0 ± 14.2</td>
</tr>
<tr>
<td>Progress Note</td>
<td>1406.6 ± 112.9</td>
<td>1993.2 ± 71.9</td>
<td>2979.6 ± 351.3</td>
<td>4842.1 ± 69.8</td>
</tr>
<tr>
<td>Surgery &amp; Procedure</td>
<td>1222.7 ± 80.2</td>
<td>1125.6 ± 35.1</td>
<td>1088.4 ± 135.2</td>
<td>1223.3 ± 22.6</td>
</tr>
<tr>
<td>All</td>
<td>5873 ± 276</td>
<td>6277 ± 178</td>
<td>6922 ± 741</td>
<td>9546 ± 133</td>
</tr>
</tbody>
</table>

**Table 5: Mean Note Volume, Per Note Type, in Notes Per Week**

**Input Modality Evolution:** Table 6 provides mean estimates for the five input modalities during the four periods of study. This research revealed a strong decline in medical transcription volume across all note types from 1467.7 +/- 28.9 notes per week to 165.4 +/- 8.7 notes per week, which represents an 88.7% decrease in transcription by number of documents, from pre- to post-implementation of SR. Regarding the volume of scanned documents in the EHR, this study found scanned note volume decreased by 34% from 3228.5 +/- 222.7 notes per week before implementation to 2121.2 +/- 30.8 notes per week after implementation. DD+KB note volume more than tripled (236%) increase from 484.2 +/- 32.6 notes per week during the pilot to 1622.6 +/- 36.3 notes per week, while DD+SR increased 561% (from 713.2 +/- 53.5 notes per week during the pilot to 4726.9 +/- 85.7 notes per week).
Table 6: Mean Note Volume per Input Modality

<table>
<thead>
<tr>
<th>Input Modality</th>
<th>Before Mean ± Std Err</th>
<th>Pilot Mean ± Std Err</th>
<th>Transition Mean ± Std Err</th>
<th>After Mean ± Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD+KB</td>
<td>0.0 ± 0.0</td>
<td>484.2 ± 32.6</td>
<td>793.2 ± 114.6</td>
<td>1622.6 ± 36.3</td>
</tr>
<tr>
<td>DD+SR</td>
<td>0.0 ± 0.0</td>
<td>713.2 ± 53.5</td>
<td>2206.4 ± 294.9</td>
<td>4726.9 ± 85.7</td>
</tr>
<tr>
<td>MedTrans</td>
<td>1467.7 ± 28.9</td>
<td>1146.2 ± 35.5</td>
<td>585.2 ± 80.1</td>
<td>165.4 ± 8.7</td>
</tr>
<tr>
<td>PowerNote</td>
<td>1176.8 ± 55.3</td>
<td>1051.2 ± 37.2</td>
<td>1016.8 ± 129.6</td>
<td>942.6 ± 19.5</td>
</tr>
<tr>
<td>ScanDoc</td>
<td>3228.5 ± 222.7</td>
<td>2974.5 ± 67.8</td>
<td>2320.4 ± 272.6</td>
<td>2121.2 ± 30.8</td>
</tr>
<tr>
<td>All</td>
<td>5873 ± 276</td>
<td>6277 ± 178</td>
<td>6922 ± 741</td>
<td>9546 ± 133</td>
</tr>
</tbody>
</table>

Figure 1 displays the evolution of note volume, per input modality, measured in number of notes per week. The study periods, per definition and timeline provided, are indicated by background color: before (red), pilot (orange), transition (purple), and after (green). Local regression (LOESS) trend lines are shown per input modality in addition to the 95% confidence intervals that are indicated as shaded bands around the trend lines. DD+KB (red curve) starts in June 2014, ramps throughout the year, increases steeply during the implementation period, and continues to increase afterwards. DD+SR (gold curve) becomes the most prevalent input modality after the implementation of SR. Scanned notes (purple curve) are the most voluminous input modality before the implementation, but the volume declines from the start of the pilot and continues to steadily decline throughout the study period. Medical transcription volume (green curve) declines throughout the study period with a sharp drop during the implementation.

Input Modality Mix Evolution: Figure 2 shows the input modality mix, as a percentage, in the four study periods. Before the pilot, 55% of inpatient provider clinical documentation was scanned handwritten notes, 25% was done using medical transcription, and 20% with the use of PowerNote. These modalities changed during the pilot, with the introduction of DD+SR. During the pilot, 11.1% of note volume was created by DD+SR and 7.5% with DD+KB. During the transition period, there was further adoption of DD+SR. Post-implementation, half (49.5%) of electronic provider documentation was done with DD+SR, 17% with DD+KB, and medical transcription was greatly reduced to 1.7%. The adoption rate of electronic provider documentation (the combination of PowerNote, DD+KB, and DD+SR) increases from 20% to 77% of provider notes. The results above indicate that SR utilization evolved throughout the study period and the majority of DD note volume is captured with the use of SR.

Figure 2: Input Modality Mix Evolution
Figure 3 shows the fraction of DD that is done with SR. Volume starts around 50% at the start of the pilot, ramps up during the pilot to 65%, and jumps to a significant 74.4% adoption of SR when using DD after implementation. The steep slope of this curve during the pilot, and especially during the implementation, indicates a rapid provider adoption of SR.

Figure 3: Adoption of SR when using DD Over Time

Financial/Cost Outcomes Analysis: Another important driver for this research was to identify the outcomes of any cost savings that resulted from the implementation of SR technology. Results, demonstrated in Figure 4, show the evolution of medical transcription cost and financial savings. Medical transcription costs at VBMC started to trend down during the pilot, because of the high volume of clinical documentation done by hospitalists as early adopters of DD+SR, but researchers sought to assess the overall impact of SR on medical transcription costs and determine whether cost savings existed.

Before the pilot, the mean medical transcription expense per month was $89,400, therefore, the estimated cost at that rate for the entire study period would have been approximately $2.86 million (counterfactual). The actual observed cost, however, was approximately $1.49 million, representing a total cost savings of $1.37 million for VBMC during the study period, or a monthly cost reduction of 81.3%. The financial savings were a direct consequence of the strong decline observed in medical transcription volume as SR was adopted by providers.

Figure 4: Evolution of Medical Transcription Cost Over Time

DISCUSSION

This is the first study of its kind, to our knowledge, based on the following: First, both qualitative and quantitative methods were utilized concurrently to measure the adoption of speech recognition and its impact on provider satisfaction, documentation quality, efficiency, and medical transcription cost when used for clinical documentation within the electronic health record. Qualitative surveys were used to gauge providers’ experiences and expectations regarding documentation before and after the deployment of SR. Second, over a period of 31 months, quantitative data was collected from the EHR, transcription and SR solutions to measure adoption and cost savings. Third, a new methodology was developed and used in this study to determine which provider notes within the EHR were created with SR, to facilitate a more accurate measurement of note volume evolution per input modality. Fourth, preliminary analysis indicated there may be positive effects on satisfaction, efficiency, and workflow of documentation reviewers, beyond providers, but this will require additional analysis and review to confirm. Validated results revealed significant improvements in satisfaction, documentation quality, and efficiency among providers as a direct result of SR use. An improved provider experience correlated to an 81% reduction in monthly medical transcription costs, an increase from 20% to 77% in electronic clinical documentation adoption, and a 74% SR adoption rate.

According to Hodgson and Coiera’s review of literature, to assess the impact of SR on clinical documentation, the mean study duration was 120 days across all studies and the number of clinical documents assessed varied from 14 to more than 300,000. The quantitative segment of this study covered a period of 31 months and analyzed more than 1 million provider notes (studied six note types: History and Physical, Consults, Progress Notes, Anesthesia Notes, Surgery & Procedure Notes, and Discharge Notes; and five input modalities: Dynamic Documentation™ [DD] using the keyboard, DD with SR, Medical Transcription, PowerNote without SR, and Scanned Handwritten Notes).

Provider satisfaction was significantly improved after the introduction of SR as evidenced by the 99-point positive shift in Net Promoter Score® and 95% of providers agreeing it was a good idea to introduce SR for documenting in the medical record. This, coupled with the
studies by Alapetite, et al.\textsuperscript{7} and Lyons, et al.\textsuperscript{8} shows a definite rise in both provider expectations and experiences regarding the use of SR, further validating advances made in SR technology. Furthermore, an interesting finding in this current study is that while expectations were high before the introduction of SR, expectations were significantly exceeded after the introduction of SR, which presents an opportunity for additional review and future research.

Study findings also revealed statistically significant improvement in documentation quality and completeness required for optimal care and a borderline statistically significant reduction in time spent answering questions and clarifications from CDI, nursing, medical records, quality, and coding staff. After the implementation of SR, 81% of providers reported improvement in the quality and completeness of clinical documentation, while 60% of providers reported spending less time answering questions and clarifications. Much of the previous research for SR quality and completeness focused heavily on comparing errors or recognition rates versus the quality and completeness provided by supporting clinical narrative capture, reduced time answering questions and clarifications, and its impact to quality and efficiency.\textsuperscript{9,10} There is well-documented feedback from professional organizations to preserve the clinical narrative necessary for quality care,\textsuperscript{7,5} which presents a key opportunity to expand upon the research findings in this study and others.

To support recommendations proposed by professional organizations,\textsuperscript{2,9} documentation should support clinicians’ cognitive needs and align with, but not impede, the care team’s workflow and delivery of care. After the implementation of SR, 84% of providers reported very significant workflow improvements. This greatly exceeded expectations reported prior to SR implementation and showed remarkable progress. There is limited research focused on the impact SR has on provider and/or care team workflow,\textsuperscript{9,10} which presents an opportunity to expand upon this study during future research.

Prior studies by Alapetite, et al.\textsuperscript{7} and Lyons, et al.\textsuperscript{8} indicated that 3% and 51% of participants, respectively, reported time savings using SR. This current study revealed that 57% of participants reported decreased time spent documenting the patient encounter when using SR, but was not statistically different from the Lyons, et al. study. These findings are also consistent with the Vogel, et al.\textsuperscript{11} study. Time savings, when using SR, can be challenging to compare and quantify due to the variation in user utilization of speech-enabled Smart Templates and auto-texts, differences among specialties, and variations in note type utilized. Yet, this presents an interesting future research opportunity regarding implementation best practices.

Post-implementation, the adoption rate of electronic provider documentation increased significantly from 20% to 77% of notes, while medical transcription was greatly reduced to 1.7%. After implementation, given the choice between SR or keyboard and mouse, providers created 74% of note volume with SR when using Dynamic Documentation. This is consistent with survey findings regarding provider satisfaction with SR. These observations could be attributed to many factors, including but not limited to effective integration, implementation, training, SR accuracy, leadership buy-in, and provider satisfaction with the new solution. Furthermore, these findings suggest that SR is indeed a factor in the adoption rate of electronic provider documentation and presents an interesting future research opportunity regarding implementation best practices. While not completed for this study, the opportunity exists to augment this study, and for future research studies, to correlate provider utilization data by input method and per note type to efficiency, workflow, quality, and provider satisfaction. Another interesting research question is the adoption rate of electronic provider documentation without SR compared to with SR.

Preliminary analysis indicated, though not presented in this paper, there may be positive effects on satisfaction, efficiency, and workflow of other documentation reviewers beyond providers. This requires additional analysis to confirm, presenting an opportunity to expand this study and future studies.

Hodgson and Coiera\textsuperscript{10} identified some form of economic evaluation in seven previous clinical documentation SR studies. Of no surprise, this current study denoted a significant decline of 81.3% per month in medical transcription costs, which translated to a total cost savings of $1.37 million at Vassar Brothers Medical Center during the study period. Furthermore, the extent and speed of the decline in medical transcription costs observed in the study, within only a few months, is noteworthy. These observations could be attributed to many factors, including but not limited to effective integration; implementation; training, which leveraged lessons learned from the pilot; leadership buy-in; and provider satisfaction with the new SR solution. A hard return on investment is difficult to calculate due to the variability involved. Variables, among others, can include the following: the impact on timely care transitions, discharges and care quality, resulting from immediate access of speech-enabled provider notes that can affect clinical and financial outcomes; financial impact from the potential reduction of delays for coding and billing through timely access to complete encounter information; efficiency gains of non-providers as a result of less time spent seeking or clarifying information, which was validated in this study as providers felt they spent less time answering questions and clarifications; and potential reduction of provider burnout and turnover rates through improved satisfaction.

LIMITATIONS

Several factors could hinder this study from being generalized to other clinical settings or information systems. Other EHR and SR systems might differ in performance and different approaches to integrating and implementing the two could also lead to varying results. Furthermore, this study was not designed to evaluate clinical effectiveness nor the direct impact on clinical outcomes of one input modality versus another. Post-implementation surveys spanned many months, potentially impacting provider perceptions of SR because of communications with and/or witnessing other providers’ use of SR, which may have created unintended bias in survey results. We were unable to collect note length information to measure the effect of SR input modality on note length. The methodology to interlink EHR notes with SR is binary, and, as such, does not provide measurement as to how much SR was used as opposed to keyboard and mouse. No time-motion studies were done with providers, which may have provided greater detail around the actual decrease in documentation time, reduced clarifications, and improved documentation quality when using SR versus other input modalities.

CONCLUSIONS

For many organizations, the search for the most efficient method of capturing clinical documentation, optimizing the EHR, enhancing usability and improving provider satisfaction remains a strategic priority. Given the substantial cost of documentation, in terms of clinician time and varying input modalities, the foundational act of interacting with an electronic health record was the primary focus of this study. By linking provider notes within the EHR to SR, and evaluating both qualitative and quantitative data inputs concurrently, this study created a baseline for the impact of SR use on provider adoption, satisfaction, documentation quality, efficiency, and cost. This new methodology, and
its associated findings, have contributed to the current field of research regarding best practices for integrating and implementing SR within the EHR.

However, as SR technologies continue to extend into different healthcare settings and advance to include more robust narrative and clinical intelligence capabilities, additional studies are needed to augment this research. Furthermore, the healthcare field should evolve the understanding of the ways SR will support providers and downstream documentation users within their workflow, and how improved documentation translates to better clinical and financial outcomes.

ACKNOWLEDGEMENTS

Special thanks to the following for supporting and/or participating in interviews, data collection, analyses, and/or project leadership: Ray Pankuch, Roxanne Parrella, Barbara McCrerey and staff, Julie Primavera and staff, Beth Falder, Lore Bogolin and staff, Mickey O’Neill and staff, Carol Wilson, Dawn Louey, Kelly Phibbs, ND, Adem Arslani, Melinda Heady, Justin Riseley, Heather Klemm, Stephanie Barr McDonald, Mercedes Wilson, John Ringel, Kelly DeQuatro, Emraan Khatri, Linda O’Grady, Francis Ganong, Beth DeLisi-Baird, and Molly Clark.

1Health Quest, Lagrangeville, New York; 2Nuance Communications, Inc., Burlington, Massachusetts

COMPETING INTERESTS

The authors indicated with the number 2 are employed by Nuance Communications Inc., a leading provider of speech recognition solutions.

REFERENCES


Identifying Supplement Use Within Clinical Notes: An Application of Natural Language Processing

Vivekanand Sharma, PhD, Indra Neil Sarkar, PhD, MLIS
Center for Biomedical Informatics, Brown University, Providence, Rhode Island

Abstract
Recent statistics indicate that the use of dietary supplements has increased over the years. Although being popular among consumers who use them for a variety of reasons, there have been limited clinical data-driven studies of the impact of dietary supplements on health outcomes. Challenges that impede such analyses in a comprehensive manner include either the sequestered nature of such data or their embedding within biomedical and clinical text. This study explored the feasibility to uncover patterns in the use of supplements, focusing on vitamin use among patients diagnosed with mental illness within patient records from the MIMIC-III database. The relevance of vitamin(s) was calculated at different levels of granularity and compared with association identified from Dietary Supplement Subset of MEDLINE. The results reveal insights into vitamin use for specific mental health related diagnosis and highlight challenges with identifying supplement information from clinical sources.

Introduction
The estimated sales of dietary supplements in the United States totaled $36.7 billion in 2014, out of which $14.3 billion (38.96%) included vitamin and mineral containing supplements. According to the 2016 annual survey on dietary supplements conducted on behalf of Council for Responsible Nutrition (CRN), about 97% of dietary supplement users take vitamins or minerals. Despite debate on their potential benefits, the choice to use vitamins may be guided by reasons including: overall health and wellness, enhance energy, fill in nutrient gaps, and immune health. The research domain focused on studying dietary supplements is fraught with challenges that include inadequate data resources, lack of suitable means to measure health outcomes, and rudimentary quantification of risks and benefits. In order to aid data-driven methodologies to gain insight about the use and impact of supplement use, acquisition of data about supplement usage from existing biomedical and clinical knowledge sources may be essential.

Patients records provide a rich source of information that can allow for analysis of real-world effectiveness and use patterns. They encompass reports of symptoms, findings from examinations, as well as diagnostic tests, diagnoses, prescribed drugs, and other interventions, along with additional information on facets such as family, social, and behavioral aspects. Such data are present in health records as either structured or unstructured fields. Patient records have been used to study patterns and trends in use of medications and their respective health outcomes. Although rich in information content, the cumulative growth of data and presence of free text elements poses significant challenges in attaining meaningful information. Automated acquisition of clinically relevant knowledge from health records may provide ways to address such issues allowing analyses and dispersal of insights in a timely and high-throughput manner. Approaches leveraging Natural Language Processing (NLP) pipelines for identification of required information from free text fields has shown promise (e.g., tobacco use information from social history and clinical notes).

Concurrently, the evolving clinically relevant knowledge with new discoveries in biomedical domain has the potential to alter clinical practice. However, such knowledge is buried in vast amounts of biomedical literature (e.g., articles indexed within MEDLINE). The insights gained in a timely manner from rich sources of information such as patient records or biomedical literature have the potential to provide the groundwork for decision support, quality assurance, and clinical informational needs. Text mining provides essential ways to address the issue of extracting relevant information from larger corpora in an automated manner. One of the primary tasks in development of such pipelines is recognition of biomedical entities embedded within the text. Leveraging the NLP based recognition of entities, several studies have attempted to identify related entities (e.g., drug-disease and gene-disease) using co-occurring concepts. Such studies have also shown potential in identification of herb-disease associations to cater specific informational needs. Systems such as SemRep and BioMedLEE have also been used for integrating co-occurrence based systems with semantic relation extraction system for literature-based discovery. The existing
NLP systems to recognize biomedical entities and relations have relied on rule-based or machine learning based approaches. More recent advancements in this area have leveraged deep learning for popular NLP tasks.

Significant advances have been made in acquisition and analyses of data related to the use of prescription drugs and associated health outcomes. However, such advancements have not yet been leveraged to its fullest potential in the domain of Natural Health Products and Supplements (NHP&S). The primary hurdle in conducting analyses of health outcomes associated with NHP&S results from data being either sequestered in domain specific sources lacking resources for interoperability or embedded within the general biomedical and clinical data sources. There are myriad challenges surrounding the tools and techniques to study NHP&S, including the lack of domain-specific vocabularies and inadequate term coverage by conventional biomedical resources. There is a need for development of tools, techniques, and resources for supporting the needs for making the data available in a standardized format that could facilitate the subsequent analyses.

Collating information on efficacy and safety of supplements may equip patient and healthcare providers with important knowledge to make informed choices, especially in light of debatable effects on health outcomes. This has been observed in case of existing understanding of the connection between dietary supplements and mental illness. With the introduction of psychiatric medications in the 1950s there has been a decline in interest in studies related to relevance of dietary nutrients for mental health. Among other supplements, the use of vitamins and minerals is more popular. While the use of this subset of supplements is popular, the pattern of use with regards to specific ailments requires further investigation. Studies have shown that low levels of vitamins can be related with mental illness, for example, association between low levels of folic acid with depressive symptomatology and poor response to antidepressant medications. Vitamin B6, B8, and B12 have been shown to reduce psychiatric symptoms and reduce the duration of illness. There is evidence in scientific literature that reflect the use of specific vitamins for treating specific psychiatric symptoms (e.g., use of Vitamin B1 to treat anxiety disorders including symptoms such as insomnia, nightmares, anorexia, nausea and vomiting). Similarly, it has been proposed that a subset of schizophrenia patients can respond to niacin augmentation therapy better than others. Such evidence suggest the utility in studying patterns of vitamin use and efficacy at different levels of diagnostic granularity.

This study examines the feasibility of mining relevant information about supplement use from patient records and provides a comparison with biomedical literature, focusing on a case study of vitamin use among patients diagnosed with mental illness. Associations related to vitamins were specifically calculated with respect to overall use of other supplements among patients with mental illness. A custom thesaurus integrated with an NLP tool was used to process the text for identification of NHP&S mentions. The feasibility and relevance of identified associations were examined by leveraging the granularity of International Classification of Diseases codes, Version 9, Clinical Modification (ICD-9-CM) and Clinical Classifications Software (CCS) Single-Level Diagnoses categories (CCS category). The results from this case study provide insight into potential improvements needed for identification and mapping approaches to support study of supplement use within clinical contexts. The results also reveal some early perspectives on vitamin use for specific diagnoses within the larger category of mental disorders and suggest that this area of research warrants a more in-depth clinical data-driven evaluation of health outcomes.

Materials and Methods

The primary goal of this study was to provide a comparative view outlining the use of vitamins with respect to other NHP&S among patients with diagnoses falling within the ICD-9-CM category of “Mental Disorders (290-319)”. A custom thesaurus of NHP&S terms integrated with biomedical NLP tool MetaMap (built using the MetaMap Data File Builder suite) was used to facilitate extraction of relevant information from the data sources of interest. Standard MetaMap was used to identify diagnoses terms from MEDLINE text and further mapped to ICD-9-CM. Associations between NHP&S terms and mental health related diagnoses terms were calculated at two levels: (1) ICD-9-CM codes; (2) CCS Single-Level Diagnoses categories (“CCS category”). A general overview of the approach is depicted in Figure 1.

Data sources and filters used for the study

Data from two sources were considered for this study: De-identified patient records from the Medical Information Mart for Intensive Care III (MIMIC-III) database and biomedical literature indexed in MEDLINE.

MIMIC-III: Provided by the Beth Israel Deaconess Medical Center, MIMIC-III is a publicly available repository containing de-identified data from over 40,000 critical care patients ranging from year 2001 to 2012. Such a resource can be leveraged to support studies including epidemiology, clinical decision-rule improvement, and
The data contained within MIMIC-III spans demographics, vital sign measurements, laboratory tests, procedures, medications, and notes and reports provided by the healthcare provider. These data elements are organized into several tables in relational form. For the purpose of this study, two of the tables were of interest: (1) DIAGNOSES_ICD: Provides a list of diagnoses for a given patient identifier coded using the ICD-9-CM; (2) NOTEEVENTS: Contains notes from providers in several categories such as discharge summary, nursing, nutrition, physician, and radiology. Focusing on diagnoses related to mental health disorders, a SQL query was used to identify data (diagnoses and text notes) associated with patients with at least one diagnosis from the ICD-9-CM category, “Mental Disorders (290-319)”. 

**Figure 1: Study Overview.** Patient records from associated with ICD-9-CM category “Mental Disorders” were processed to identify mentions of dietary supplements. Corresponding articles related to the identified ICD-9-CM codes were identified from MEDLINE (Dietary Supplement Subset). Associations between supplements and diagnoses (ICD-9-CM codes or CCS Categories) were calculated and results were presented using vitamin use as a case study.

**MEDLINE:** Provided by the National Library of Medicine (NLM), MEDLINE is one of the primary sources of biomedical literature and references. It contains more than 24 million references from journals selected based on a comprehensive selection process. The article entries within MEDLINE are indexed using a controlled vocabulary, the Medical Subject Headings (MeSH). Focused specifically on articles related to the use of NHP&S by patients diagnosed with mental health issues, the search relied on a subset of literature related to dietary supplements. This “Dietary Supplement Subset” is provided by the partnership of the Office of Dietary Supplements and the NLM to restrict the search results to articles related to a broad spectrum of literature including vitamins, minerals, phytochemicals, ergogenic, botanical, and herbal supplements in human nutrition and animal models. Using this subset as a primary filter, an additional filter for mental health disorders was applied using MeSH descriptor “Mental Disorders (F03)” (Query: “Mental Disorders”[mh] AND dietsuppl[sb]). An additional article set where mental disorders was the central theme was also identified (Query: “Mental Disorders”[majr] AND dietsuppl[sb]) and retained for comparison.

**Custom thesaurus of NHP&S terms.** A custom thesaurus was used for this study for domain specific information acquisition (created as a part of previous study). This thesaurus was built to incorporate NHP&S terms from sources that offer reliable and comprehensive coverage of terms, synonyms, and variants. The selection of terms was based on sources that aim to provide evidence-based information to healthcare providers and general public. The primary selection of terms was from six sources: LNHPD, DSLD, SRS-UNII, RxList, Natural Medicines, and Medscape. Similar strings, synonyms, and variants across different sources were grouped together into concepts and assigned unique identifiers. This dataset was organized into tables for use with MetaMap Data File Builder suite (2016) to create a custom thesaurus that could be used with the NLP tool MetaMap (“Custom MetaMap”).
**Processing of MIMIC notes.** The notes identified from MIMIC-III database table NOTEEVENTS were stored as separate text files for each patient. These text files were processed using Custom MetaMap and the machine output was parsed using a Julia program retaining mappings with a perfect score of 1000.

**Processing of MEDLINE abstracts.** The titles and abstracts were extracted from the MEDLINE subset in XML format. The resulting set of text was processed separately using MetaMap Java API with: (1) Custom MetaMap: To identify NHP&S mappings (with score of 1000); (2) Default MetaMap: To identify mappings to diagnoses terms. The mappings from text to ICD-9-CM codes were performed using a two-fold approach: (a) Direct Mapping: The default MetaMap processing was restricted to ICD-9-CM (~q→ICD9CM). The UMLS CUIs identified with perfect mapping scores were used to directly query UMLS MRCONSO table to retain a corresponding ICD-9-CM code; and (b) Inferred Mapping: The text was processed using default MetaMap and the output was parsed to retain UMLS CUIs (with mapping score of 1000) of semantic types belonging to the UMLS semantic group “Disorders”. The UMLS CUIs were queried against UMLS MRCONSO to identify direct mappings. If no direct mappings were found, a recursive search query (five iterations) was performed on UMLS MRREL table (including relationship type ‘isa’) to identify mappings to ICD-9-CM. This search approach followed UMLS CUIs within MRREL table to identify entities that are related by relationship type ‘isa’ until a CUI was identified that could be directly mapped to ICD-9-CM. The identification of ‘isa’ related CUIs and checking for mapping to ICD-9-CM codes was pursued up to five iterations.

Following the processing of text and identifying NHP&S terms and corresponding ICD-9-CM codes, those records that had ICD-9-CM codes identified from list of diagnoses from MIMIC dataset within the ICD category “Mental Disorders (290-319)” were retained. For all ICD-9-CM codes, their corresponding CCS Single-Level Diagnoses categories were identified and used for further analyses. The CCS groups more than 14,000 ICD-9-CM diagnosis codes into a smaller number of clinically meaningful categories. The categories include classifications for Mental Health and Substance Abuse (CCS-MHSA)\(^{32}\).

**Associations among NHP&S terms and ICD9CM diagnoses.** For both datasets, associations between NHP&S terms and diagnoses terms (both individual ICD-9-CM codes and CCS categories) were calculated using Odds Ratio (OR) with a threshold of at least five or more subjects with a given ICD-9-CM diagnosis using a particular vitamin. Associations with OR greater than one, having 95% Confidence Interval (95% CI) that did not include one were retained for further comparison between the MIMIC and MEDLINE datasets.

**Results**

**Top ICD-9-CM codes with respect to vitamin use.** A total of 13,400 subjects were identified from MIMIC-III data having at least one diagnoses falling in the “Mental Disorder (290-319)” category. Out of 13,400 subjects, 3248 (24.24%) use one or more vitamins. The demographics of study population and vitamin users is provided in Table 1.

<table>
<thead>
<tr>
<th>Table 1: Demographics of subject population</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sex</strong></td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>Female</td>
</tr>
<tr>
<td>&gt; 5</td>
</tr>
<tr>
<td>05-19</td>
</tr>
<tr>
<td>20-34</td>
</tr>
<tr>
<td>35-49</td>
</tr>
<tr>
<td>50-64</td>
</tr>
<tr>
<td>≥65</td>
</tr>
<tr>
<td><strong>Age Groups</strong></td>
</tr>
<tr>
<td><strong>Ethnicity (Top Five)</strong></td>
</tr>
<tr>
<td>White</td>
</tr>
<tr>
<td>Black</td>
</tr>
<tr>
<td>Unknown</td>
</tr>
<tr>
<td>Hispanic</td>
</tr>
<tr>
<td>Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Vitamin use by subjects categorized by ICD-9-CM (threshold 10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ICD-9-CM</strong></td>
</tr>
<tr>
<td>311</td>
</tr>
<tr>
<td>305.1</td>
</tr>
<tr>
<td>291.81</td>
</tr>
<tr>
<td>293.0</td>
</tr>
<tr>
<td>303.91</td>
</tr>
<tr>
<td>300.00</td>
</tr>
</tbody>
</table>
The top ICD-9-CM diagnoses codes for which vitamin use was prevalent (with a threshold of 10% of total vitamin users) were: (1) Depressive disorder NEC; (2) Tobacco use disorder; (3) Alcohol withdrawal; (4) Acute delirium; (5) Alcohol dependence; and (6) Anxiety state NOS. The count and their respective percentages are listed in Table 2.

**Distribution across CCS Single-Level Diagnoses Categories.** The ICD-9-CM codes identified from subjects using one or more vitamins were mapped to 14 CCS categories. The top CCS categories (threshold of 10%) with higher number of vitamin users were (CCS category code in parenthesis): (1) Alcohol-related disorders (660); (2) Mood disorders (657); (3) Delirium, dementia, and amnestic and other cognitive disorders (653); (4) Screening and history of mental health and substance abuse codes (663); (5) Substance-related disorders (661); and (6) Anxiety disorders (651). The distribution of the total of 3244 subjects using vitamins with at least one diagnoses in mental disorder category is listed in Table 3.

<table>
<thead>
<tr>
<th>CCS Code</th>
<th>CCS Category Name</th>
<th>Subjects</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>650</td>
<td>Adjustment disorders</td>
<td>35</td>
<td>01.08</td>
</tr>
<tr>
<td>651</td>
<td>Anxiety disorders</td>
<td>433</td>
<td>13.33</td>
</tr>
<tr>
<td>652</td>
<td>Attention-deficit/conduct/disruptive behavior disorders</td>
<td>25</td>
<td>00.77</td>
</tr>
<tr>
<td>653</td>
<td>Delirium, dementia, and amnestic and other cognitive disorders</td>
<td>853</td>
<td>26.26</td>
</tr>
<tr>
<td>654</td>
<td>Developmental disorders</td>
<td>96</td>
<td>02.96</td>
</tr>
<tr>
<td>655</td>
<td>Disorders usually diagnosed in infancy, childhood, or adolescence</td>
<td>5</td>
<td>00.15</td>
</tr>
<tr>
<td>656</td>
<td>Impulse control disorders, NEC</td>
<td>2</td>
<td>00.06</td>
</tr>
<tr>
<td>657</td>
<td>Mood disorders</td>
<td>1238</td>
<td>38.12</td>
</tr>
<tr>
<td>658</td>
<td>Personality disorders</td>
<td>38</td>
<td>01.17</td>
</tr>
<tr>
<td>659</td>
<td>Schizophrenia and other psychotic disorders</td>
<td>148</td>
<td>04.56</td>
</tr>
<tr>
<td>660</td>
<td>Alcohol-related disorders</td>
<td>1271</td>
<td>39.13</td>
</tr>
<tr>
<td>661</td>
<td>Substance-related disorders</td>
<td>469</td>
<td>14.44</td>
</tr>
<tr>
<td>663</td>
<td>Screening and history of mental health and substance abuse codes</td>
<td>640</td>
<td>19.70</td>
</tr>
<tr>
<td>670</td>
<td>Miscellaneous mental health disorders</td>
<td>65</td>
<td>20.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CCS Code</th>
<th>ICD-9-CM</th>
<th>ICD-9-CM description</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>660</td>
<td>291.81</td>
<td>Alcohol withdrawal</td>
<td>404</td>
<td>12.44</td>
</tr>
<tr>
<td>303.91</td>
<td>Alcohol dependence</td>
<td>358</td>
<td>11.02</td>
<td></td>
</tr>
<tr>
<td>657</td>
<td>311</td>
<td>Depressive disorder NEC</td>
<td>815</td>
<td>25.09</td>
</tr>
<tr>
<td>653</td>
<td>293.0</td>
<td>Acute delirium</td>
<td>387</td>
<td>11.92</td>
</tr>
<tr>
<td>663</td>
<td>305.1</td>
<td>Tobacco use disorder</td>
<td>640</td>
<td>19.70</td>
</tr>
<tr>
<td>661</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>651</td>
<td>300.00</td>
<td>Anxiety state NOS</td>
<td>336</td>
<td>10.35</td>
</tr>
</tbody>
</table>

**ICD-9-CM codes within CCS categories.** Keeping a threshold proportion of 10%, the top ICD-9-CM diagnoses among vitamin users for CCS category “Alcohol-related disorders (660)” was Alcohol withdrawal (291.81) and Alcohol dependence (303.91). Similarly, for “Mood disorders (657)”, the top ICD-9-CM code was Depressive disorder NEC (311). For CCS category “Delirium, dementia, and amnestic and other cognitive disorders (653)” the top ICD-9-CM code was Acute delirium (293.0). For “Screening and history of mental health and substance abuse codes (663)” the top ICD9CM code was Tobacco use disorder (3051) and for “Anxiety disorders (651)” the ICD9CM code was Anxiety state NOS (300.00). A summary of the proportions of top ICD-9-CM diagnoses within top CCS categories is listed in Table 4. A list of vitamins significantly associated (as identified from MIMIC-III and MEDLINE dataset) with the top ICD-9-CM diagnoses codes mentioned above is presented in Table 5.

**Vitamin use associations among CCS categories.** Significant associations of vitamin use and ICD-9-CM diagnoses were identified for 11 out of the initially identified 14 CCS categories. These associations were either from MIMIC-III dataset or MEDLINE dataset (using direct UMLS CUI to ICD-9-CM mapping or inferred UMLS CUI to ICD-9-CM mapping). For the MEDLINE dataset, a higher number of associations were identified when the inferred
mapping (via UMLS MRREL relationship type ‘isa’) was used. Common associations among MIMIC-III and MEDLINE dataset were identified for only three CCS categories: (1) Delirium, dementia, and amnestic and other cognitive disorders (653): Vitamin E and B12; (2) Alcohol-related disorders (660): Vitamin B1; and (3) Miscellaneous mental health disorders (670): Vitamin D. The results from identification of associations using odds ratio across different CCS categories for MIMIC and MEDLINE dataset is summarized in Table 6.

Table 5: Vitamin associated with ICD-9-CM codes from Table 4

<table>
<thead>
<tr>
<th>ICD-9-CM</th>
<th>ICD-9-CM description</th>
<th>MIMIC-III</th>
<th>MEDLINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>291.81</td>
<td>Alcohol withdrawal</td>
<td>Multivitamin: 9.71 3.31-28.49 p &lt; 0.0001</td>
<td>Vitamin B1: 1.48 2.39-16.02 p = 0.0002</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B1: 6.42 5.50-7.49 p &lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B9: 5.22 4.50-6.05 p &lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B12: 3.12 1.73-5.63 p = 0.0002</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B12: 2.69 1.33-5.43 p = 0.0057</td>
<td></td>
</tr>
<tr>
<td>303.91</td>
<td>Alcohol dependence</td>
<td>Vitamin B1: 7.10 4.47-11.29 p &lt; 0.0001</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B9: 4.41 3.77-5.16 p &lt; 0.0001</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multivitamin: 3.73 2.14-6.51 p &lt; 0.0001</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin K: 3.69 2.96-4.60 p &lt; 0.0001</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin K: 2.41 1.15-5.04 p = 0.0194</td>
<td>NA</td>
</tr>
<tr>
<td>311</td>
<td>Depressive disorder</td>
<td>Vitamin D2: 2.08 1.56-2.78 p &lt; 0.0001</td>
<td>Vitamin B9: 1.60 1.15-2.22 p = 0.0050</td>
</tr>
<tr>
<td>NEC</td>
<td></td>
<td>Vitamin D: 1.54 1.32-1.80 p &lt; 0.0001</td>
<td>Vitamin D1: 1.51 1.02-2.23 p = 0.0370</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B12: 1.41 1.07-1.85 p = 0.0143</td>
<td></td>
</tr>
<tr>
<td>293.0</td>
<td>Acute delirium</td>
<td>Vitamin K: 1.92 1.43-2.58 p &lt; 0.0001</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B12: 1.87 1.03-3.41 p = 0.0411</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin D: 1.56 1.03-2.36 p = 0.0339</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin D: 1.41 1.14-1.74 p = 0.0014</td>
<td>NA</td>
</tr>
<tr>
<td>305.1</td>
<td>Tobacco use disorder</td>
<td>Vitamin D: 1.96 1.27-3.03 p = 0.0026</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin D2: 1.52 1.01-2.30 p = 0.0469</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin D3: 1.33 1.07-1.65 p = 0.0088</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Vitamin B12: 1.48 1.04-2.11 p = 0.0286</td>
<td>NA</td>
</tr>
</tbody>
</table>

1MEDLINE query "Mental Disorders[mh] AND dietsuppl[sb]; 2MEDLINE query "Mental Disorders[majr] AND dietsuppl[sb]; The OR scores provided in table indicate the results obtained from "Mental Disorders[mh] AND dietsuppl[sb]

Discussion

There is a continuous increase in spending on and use of dietary supplements including vitamins which may be unsupervised or prescribed by healthcare provider. In light of such growth in use of supplements, it becomes imperative to systematically study and analyze the associated health outcomes including both aspects whether there is actually any benefit and if there are any risks involved. Towards supporting such ventures to examine the efficacy and safety, extracting information from existing biomedical literature and health records may provide valuable insights. Although the importance of such sources is acknowledged, the magnitude of available data is increasing. Automated methods are therefore increasingly required for efficient cataloguing, indexing and retrieval of relevant information. Such pipelines will be essential to cater the needs of clinicians and patients to make informed choices. Though a case study of vitamin supplement use compared to other supplements among patients with mental illness, patterns in use and associations with more specific (granular) diagnoses was examined here from biomedical literature indexed in MEDLINE and patient records in MIMIC-III database. Using ICD-9-CM codes and categorization in conjunction to the CCS categories associated with mental disorders, both a general and granular account is illustrated.

Within MIMIC-III, this study indicates that a sizeable portion (24.24%) of patients with diagnoses falling in mental disorder category use a range of vitamins. The use by specific diagnosis (on the basis of ICD-9-CM codes) show higher use by patients with depressive disorder, tobacco use disorder, alcohol-related disorders, acute delirium and anxiety compared to other diagnoses (Table 2). A relatively broader picture is provided in Table 3 from where CCS categories with comparatively higher proportions of use is reflected. The highest proportion of use is by patients with diagnosis in the category of “Alcohol-related disorders” followed by “Mood disorders”, “Delirium, dementia,
and amnestic and other cognitive disorders”, “Screening and history of mental health and substance abuse codes” and “Anxiety disorders” among others. More specific ICD-9-CM codes from the above mentioned categories along

Table 6: Associations across different CCS categories for MIMIC and MEDLINE dataset as reflected from Odds Ratio (OR) and 95% Confidence Interval (CI)

<table>
<thead>
<tr>
<th>CCS</th>
<th>MIMIC</th>
<th>MEDLINE-DIRECT ICD-9-CM MAPPING</th>
<th>MEDLINE-INFERRED ICD-9-CM MAPPING</th>
<th>COMMON</th>
</tr>
</thead>
<tbody>
<tr>
<td>650</td>
<td>Vitamin D2</td>
<td>3.96</td>
<td>1.82-8.59</td>
<td>NA</td>
</tr>
<tr>
<td>651</td>
<td>Vitamin D2</td>
<td>1.58</td>
<td>1.04-2.41</td>
<td>NA</td>
</tr>
<tr>
<td>653</td>
<td>Vitamin E&lt;sup&gt;1,2&lt;/sup&gt;</td>
<td>3.22</td>
<td>1.73-6.00</td>
<td>NA</td>
</tr>
<tr>
<td>654</td>
<td>Vitamin D3</td>
<td>1.89</td>
<td>1.35-2.65</td>
<td>Vitamin B12&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>655</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>657</td>
<td>Vitamin D2</td>
<td>1.97</td>
<td>1.45-2.67</td>
<td>NA</td>
</tr>
<tr>
<td>658</td>
<td>Multivitamins</td>
<td>2.02</td>
<td>1.08-3.77</td>
<td>NA</td>
</tr>
<tr>
<td>659</td>
<td>Multivitamins</td>
<td>1.42</td>
<td>1.02-1.98</td>
<td>Vitamin B3&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>660</td>
<td>Multivitamins</td>
<td>3.06</td>
<td>2.41-3.88</td>
<td>Vitamin B1&lt;sup&gt;1,2&lt;/sup&gt;</td>
</tr>
<tr>
<td>661</td>
<td>Vitamin B1</td>
<td>1.54</td>
<td>1.31-1.81</td>
<td>NA</td>
</tr>
<tr>
<td>670</td>
<td>Vitamin D&lt;sup&gt;1,2&lt;/sup&gt;</td>
<td>1.94</td>
<td>1.08-3.50</td>
<td>Vitamin D&lt;sup&gt;1,2&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>1</sup>MEDLINE query "Mental Disorders[mh] AND dietsuppl[sh];<sup>2</sup>MEDLINE query "Mental Disorders[majr] AND dietsuppl[sh]; The OR scores provided in table indicate the results obtained from "Mental Disorders[mh] AND dietsuppl[sh]; p-values calculated for all associations was < 0.05.}

with the proportions of vitamin use is provided in Table 4. Associations of several vitamins was identified from MIMIC-III dataset (Table 5) for the identified ICD-9-CM codes. Among other vitamins, Vitamin B1 shows significant association for alcohol withdrawal diagnosis. This observation was corroborated by analyses from MEDLINE data. Vitamin B1 (Thiamine) has been studied extensively for treatment of patients with alcohol dependence<sup>18</sup>. The MEDLINE query for this association ("Alcohol-Related Disorders[mh] AND "thiamine"") results in 288 articles. A limitation that needs to be considered in the inference pipeline is that the mapping of MEDLINE text with ICD-9-CM codes is not a perfect art. As reflected from Table 6, additional associations can be identified from MEDLINE when using an inferred mapping approach. Unlike the clinical notes and diagnoses mentions, the mentions of vitamin use for a given disorder may not be very specific, resulting in missing of some associations. Additionally, the possibility of information loss as a result of incomplete collection of articles indexed as “Dietary Supplement Subset” cannot be ignored. Significant association of Vitamin B9 for depressive disorder was identified from MEDLINE which was missing from MIMIC-III dataset. It has been shown that Vitamin B9 exerts antidepressant effects by upregulating brain-derived neurotrophic factor and glutamate receptor 1 expression.
in brain. The use of Vitamin E and B12 for the CCS category “Delirium, dementia, and amnestic and other cognitive disorders (653)” as reflected from MIMIC-III data is corroborated by MEDLINE. There is considerable amount of literature that reflects Vitamin B3 association with schizophrenia treatment. The analysis in this study from MEDLINE dataset also reflects significant association between Vitamin B3 and CCS category “Schizophrenia and other psychotic disorders (659)”. However, such association was not recovered from MIMIC dataset. With respect to specific Vitamin D associations, a potential confounding factor might be the diagnostic test referred to by the physician which may have been picked up by the NLP system. The patterns of use of supplements in general may not completely reflect and align with evidence-based associations identified due to mostly lack of physician supervised nature of intake.

This study highlights several challenges and opportunities in mining patient data and biomedical literature for understanding the patterns of use of supplements. One specific aspect is having a supplement thesaurus attached to an NLP system that can provide full coverage and efficient mapping of supplement terms. Inadequate coverage of supplement terms has been shown to be a major issue in this area. Issues related to processing of clinical notes related to incomplete coverage of supplements has also been highlighted. Zhang et al. point out the importance of clinical notes in mining and assessment of clinical effects of supplement interventions and also convey the difficulties in documentation as a results of gap between supplement term list and standard biomedical terminologies. Additional challenges in such analyses lies in annotation of MEDLINE text with more specific diagnoses codes that may enhance the retrieval of key associations.

Although the custom NLP pipeline for identification of NHP&S terms used in this feasibility study was evaluated in a prior study with FDA Adverse Event Reporting System (FAERS) dataset (Precision: 0.93; Recall: 0.66; and F-score: 0.77), there is a need to further evaluate the utility of the system within a clinical context. Nonetheless, preliminary insights can be derived from existing studies that provided evaluation of efficacy of MetaMap on clinical notes, as demonstrated here. The focus of this study was on whether supplement information could be identified from within clinical narratives; however, future work is needed to assess the clinical validity or utility of the identified correlations. Additionally, further improvements in identification and extraction of entities of interest could be attained using recent advances in NLP (e.g., leveraging deep learning approaches). Mapping of text from MEDLINE to ICD-9-CM codes is another challenging aspect that needs further improvement and evaluation. In order to retain the granularity aspect of associations, the approach followed in this study attempted to infer mappings by identifying related UMLS concepts rather than traversing the hierarchy for consolidation of ICD-9-CM codes from same higher level category. Manual inspection of a portion of EHR data from MIMIC-III reflects that most mentions of vitamins and other supplements are included under the section heading “Medications on Admission” and some or all of them are continued at the time of discharge as indicated in the section “Discharge Medications”. Extracting and restricting this study to specific sections on medications and diagnoses at the time of admission may potentially mask the confounding factor resulting from mentions in other sections of the clinical note not indicative of supplement use by the patient. Such an approach may portray more conclusive results by reducing biases associated with ICU-only nature of MIMIC-III dataset. The interpretation of results from this preliminary does require some scrutiny; however, this was an exploratory proof of concept study that suggests further data-driven investigation into patterns of vitamin use. Future work will be aimed at exploring such patterns of use in more general EHR data from more than one sources.

Conclusion

Published studies have suggested an association between vitamins and mental health disorders; however, the benefits associated with more granular diagnoses still needs to be analyzed. Development of automated methods for mining and evaluation of such information from clinical notes and biomedical literature may be valuable in terms of providing noticeable leads. Through a case study, focused on vitamin use in the context of mental health conditions, we attempted to examine the feasibility of identifying such associations with respect to other dietary supplements at various levels of granularity. Several promising leads were highlighted that suggest potential vitamin use for mental health related diagnosis. Additionally, this feasibility study reveals the potential for adapting existing biomedical tools and resources for effective mapping and identification of supplement information from clinical text.

Acknowledgements

We thank Dr. Paul Stey, Dr. Isabel Restrepo, Ashley Lee, and Dr. Elizabeth Chen for useful discussions and providing essential technical guidance with analyzing the MIMIC-III database. This work was funded in part by
References

23. MIMIC [Internet]. [cited 2017 Sep 12]. Available from: https://mimic.physionet.org/about/mimic/
26. Fact Sheet Medical Subject Headings (MeSH®). 1999 Sep 1 [cited 2017 Sep 12]; Available from:
Word Repetition in Separate Conversations for Detecting Dementia: A Preliminary Evaluation on Data of Regular Monitoring Service
Kaoru Shinkawa, MS, Yasunori Yamada, PhD
IBM Research - Tokyo, Tokyo, Japan

Abstract
For detecting early signs of dementia, monitoring technology has been actively investigated due to the low diagnostic coverage as well as the requirement for early intervention. Although language features have been used for detecting the language dysfunctions resulting from dementia in neuropsychological tests, features that can be extracted by regular conversations remain unexplored. Here, we propose a feature to characterize the atypical repetition of words on different days which is observed in patients with dementia. We tested it on data obtained from a daily monitoring service for eight elderly people, including two who had been diagnosed with dementia. We found that our feature outperformed the existing linguistic features used in previous studies, such as vocabulary richness and repetitiveness, in terms of effect size and AUC score. The results suggest that the use of our proposed feature holds promise for improving detection performance in everyday situations such as regular monitoring.

Introduction
As the worldwide elderly population increases, the incidence of dementia and Alzheimer’s disease (AD) is becoming an increasingly serious health and social problem. Alzheimer’s Disease International estimated 7% of the world’s population over 65 years old has AD or a related dementia. It also reported that the worldwide cost of dementia may be as high as 605 billion USD a year, equivalent to 1% of the entire world’s gross domestic product. Among the participating countries, Japan is one of those facing a severe aging problem and thus its problems in this regard are very serious. The prevalence of dementia for persons 65 years or older is estimated at around 15%. The cost needed to address this problem in Japan is very high and reaches around 120 billion USD (14.5 trillion JPY). This corresponds to 3% of the entire Japanese gross domestic product. These figures have led to an increasing focus on dementia in recent years. In particular, early diagnosis and intervention has been increasingly recognized as a possible way of improving dementia care, because of recent failures in both clinical trials and laboratory work in the stages of AD. However, diagnostic coverage worldwide remains low. Even in high-income countries, only 40-50% of people afflicted with dementia have received a diagnosis. For example, Kotagal et al. suggested that 55% of people afflicted with dementia do not receive clinical cognitive evaluations in the United States. Monitoring technology capable of detecting early signs of dementia and AD in everyday situations has great potential for supporting earlier diagnosis and intervention. Although there are already several projects and services for monitoring the health of elderly persons with frequent data collection by using mobile applications, whether and how we can exploit the data collected on a daily basis to detect dementia has been largely unexplored.

To detect the health state of elderly persons afflicted with dementia in everyday situations, one of the most prominent candidates is identifying the evolution of language change over the course of dementia and AD. While memory impairment due to shrinking of the medial temporal lobe is the most typical symptom of dementia, both retrospective analyses and prospective cohort studies have shown that language problems are prevalent dating from presymptomatic periods. In addition, language dysfunctions at the time of diagnosis from pathologically proven AD patients with postmortem examination have also been reported. Ahmed et al. showed that they exhibited significant language changes such as syntactic simplification and impairments in lexico-semantic processing. On the basis of these findings, in previous computational work much attention has been paid to explore useful features that can be extracted from data gathered while participants performed neuropsychological tests by professionals such as medical doctors. They have found that the different stages of dementia and AD exhibit specific patterns of language changes in different domains such as phonetics and phonology, morphology, lexicon and semantics, syntax, and pragmatics. Recently, studies on investigating whether these language dysfunctions observed in mental illness including dementia could be extracted under conditions close to those of everyday life have started and have garnered increas-
ing attention. Although daily monitoring makes it possible to extract language features at one time, comparing language patterns on different days is another possible approach but remains largely uninvestigated.

In this study, we proposed a language feature of dementia that can be extracted by comparing language patterns on different days, and investigated whether it is useful for detecting dementia. To this end, we used conversational data obtained from a regular monitoring service for elderly people in Japan. We investigated the feature using real conversational data provided by a regular monitoring service, and showed how our proposed feature could be useful for detecting dementia.

**Related Work**

In this section, we will show related work that assisted us in aiming to determine the sort of features that can be extracted from conversations through regular monitoring. To this end, we will describe the acoustic and linguistic features in previous studies on speech data where individuals performed neuropsychological tests.

Acoustic features have been widely used for quantifying the individual’s state including emotion, stress, and neurodegenerative diseases such as depression and dementia. One reason for this is that these acoustic features are relatively easy to obtain compared with linguistic features requiring speech-to-text. In the context of detecting dementia, acoustic features have been used for analyzing verbal fluency tasks or speeded-up word-list generation. Previous studies reported that patients with dementia tend to increase their periods of silence as well as the number of pauses they use.

The short-term memory loss associated with dementia makes ordinary conversation difficult because of language dysfunctions such as word-finding and word-retrieval difficulties. These language dysfunctions have typically been characterized by using linguistic features. Picture descriptions were often used for these features and there are several large datasets for speech data in picture description tasks, such as DementiaBank. Here we will provide a brief description of linguistic features widely used in these tasks, because the linguistic features of spontaneous speech used in them could be useful for characterizing everyday conversations.

Syntactic complexity is closely associated with the incidence of dementia. For example, Kemper et al. showed that AD accelerates age-related deterioration in syntactic complexity compared with healthy controls. Syntactic complexity was measured in various ways, such as the mean length of sentences, “part-of-speech” frequency, and dependency distance. Dependency distance infers the number of intervening words between two syntactically related words in a sentence. Another category that must be considered is vocabulary richness. This category measures lexical diversity, which tends to reduce in dementia cases. It was calculated by three typical measures: type-token ratio (TTR), Brunet’s index (BI), and Honoré’s statistic (HS). TTR compares the total distinct word types ($U$) to the total word count ($N$) as $TTR=U/N$. Using the same $U$ and $N$, BI is also defined as $BI=U^{\frac{1}{N}}-0.165$. Unlike other measures related to vocabulary richness, for this measure the vocabulary richness becomes greater as BI becomes smaller. HS gives particular importance to unique vocabulary items used only once, also known as hapax legomena ($N_{uni}$). HS is defined as $HS=100\log N/(1-N_{uni}/U)$. A third feature category is repetitiveness. Some matrices measure the frequency of repeated words and phrases, and others estimate sentence similarities by calculating the cosine distance between two sentences. Another feature category, one widely used in image description tasks, is semantic density. It was calculated on the basis of “informational units” that are predefined objects or text segments that might refer to important information. For example, in the Boston Cookie Theft picture description task, information units consist of objects such as “Woman”, “Cookies”, and “Boy taking the cookie”. With the information units, semantic density can be defined as the number of information units divided by the total number of words. A number of previous studies have found that individuals with dementia tend to produce speech with lower information, defined as semantic density, than with healthy controls. Semantic density would be difficult to use in daily conversations, because it requires predefined information units.

With regard to language dysfunctions observed in conversations in everyday situations, a previous study based on family reports pointed out atypical word repetition. While this repetition was typically reported to occur in the same
conversation, it also appears in separate conversations that may be held on different days. Repetition in conversation across a number of days be assumed to represent memory impairment that prevents speakers from remembering recent conversations and degrading their social engagement ability by making them less able to expand on conversation topics. Therefore, in this study, we hypothesize that atypical repetition in conversation could be useful for detecting dementia in regular conversations. An attempt was made to infer that atypical repetition of words in a single conversation involved not only repetitiveness but also vocabulary richness because the repetitions could result in a small number of distinct words being used in a conversation. Thus, it might be possible to capture atypical repetition of words in separate conversations by using repetitiveness and/or vocabulary richness.

Feature Related to Word Repetition in Conversations on Different Days

In this study, we focused on repetition in conversation as one of the prominent behaviors of dementia. Previous studies characterized atypical repetition of words by the frequency of repeated words or similarity between the sentences in a single conversation. Although descriptive analysis of AD clinical interviews has reported that repetition may also appear in separate conversations that may be held on different days, features that can be extracted from separate conversations remain largely uninvestigated. In this study, we propose a feature that aims to measure word repetition, especially in conversations on different days as addition to the single conversations on the same day.

We first got pairs of conversational data $D_i$ and $D_j$ separated by $t$ days ($T - M \leq t \leq T + M$) days. HS$_i$ and HS$_j$ were extracted from $D_i$ and $D_j$ by calculating Honoré’s statistic (HS). HS is defined as $HS = 100 \log N / (1 - N_{uni}/U)$, where $N$ is the total number of words, $U$ is the total number of distinct word types, and $N_{uni}$ is the number of total distinct word types used only once. Next, we defined $D_{ij}$ as a combined document of $D_i$ and $D_j$ and extracted HS$_{ij}$ as a feature of repetitiveness in conversation on different days. Finally, our feature $R$ was calculated as follows:

$$R = w_i HS_{i}^{-1} + w_j HS_{j}^{-1} + w_{ij} HS_{ij}^{-1},$$

$$w_i + w_j + w_{ij} = 1.$$  

Weights $w_i$, $w_j$ and $w_{ij}$ are hyper-parameters selected by parameter optimization. $M$ was set to two days in this study.

Conversational data from a regular monitoring service

To test the proposed feature, we used daily conversation data obtained from a monitoring service for elderly people provided by Cocolomi Co., Ltd (http://cocolomi.net/). The purpose of their service is to help children to build a connection with their parent living alone by sharing the daily life information of elderly people, such as their physical condition. The communicator calls elderly people once or twice a week to have a daily conversation for about ten minutes. Each conversation is transcribed in spoken word format by the communicator and sent to the family by email (Figure 1). The spoken words of the communicator are eliminated.

The conversational data for analyzing were collected from eight Japanese people (five females and three males; age range 66-89 years, i.e., 82.37 ± 5.91 years old). Two of them were reported as suffering from dementia from the family. Table 1 shows the duration of the service, the number of report calls, the average duration of each call, and the average word length of each report. In total, 458,738 words were used for the analysis. All reports were written in Japanese. For preprocessing, we performed word segmentation, part-of-speech tagging, and word lemmatization on the conversation data. The words tagged as numerals and symbols were excluded from the analyzing data. For the preprocessing, we used the Japanese morphological analyzer MeCab.

Results

We investigated our proposed feature $R$ with conversational data obtained during the phone calls with the regular monitoring service. The feature aims to measure word repetition, especially in conversations that are separated by $t$ days in addition to a single conversations conducted within a single day.

First, we investigated whether our proposed feature could be used as a measure for discriminating dementia from controls. The discriminative power was measured by using both effect size (Cohen’s d) and area under the receiver
Figure 1: Overflow of regular monitoring service. A communicator calls an elderly customer once or twice a week, transcripts the conversations, and e-mails the transcripts to family members. In this study we analyzed the conversation transcripts this service provided.

Table 1: Participant data list.

<table>
<thead>
<tr>
<th>Status</th>
<th>Gender</th>
<th>Age</th>
<th>Start</th>
<th>End</th>
<th>No. of calls</th>
<th>Ave. call time Mean (SD) [min.]</th>
<th>Ave. word length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>F</td>
<td>75-77</td>
<td>2015 Mar</td>
<td>2017 Apr</td>
<td>75</td>
<td>11.21 (8.85)</td>
<td>395.13 (124.18)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>80-83</td>
<td>2014 Jul</td>
<td>2017 Apr</td>
<td>109</td>
<td>16.63 (4.47)</td>
<td>734.34 (195.10)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>87-89</td>
<td>2016 Jan</td>
<td>2017 May</td>
<td>104</td>
<td>11.15 (4.46)</td>
<td>418.86 (235.12)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>66-70</td>
<td>2014 Jul</td>
<td>2017 Apr</td>
<td>133</td>
<td>10.62 (2.32)</td>
<td>482.89 (118.95)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>78-81</td>
<td>2014 Dec</td>
<td>2016 Mar</td>
<td>72</td>
<td>12.06 (2.83)</td>
<td>554.69 (119.03)</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>82-85</td>
<td>2014 Nov</td>
<td>2017 Apr</td>
<td>226</td>
<td>17.75 (6.29)</td>
<td>572.12 (235.49)</td>
</tr>
<tr>
<td>Dementia</td>
<td>F</td>
<td>85-86</td>
<td>2014 Jul</td>
<td>2015 Nov</td>
<td>40</td>
<td>9.29 (2.15)</td>
<td>462.28 (204.12)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>88-88</td>
<td>2014 Jul</td>
<td>2014 Nov</td>
<td>13</td>
<td>7.77 (1.72)</td>
<td>277.94 (151.47)</td>
</tr>
</tbody>
</table>

operating characteristic curve (AUC-ROC). For Cohen’s d, the 0.8 effect size can be assumed to be large, while the 0.5 effect size is medium and the 0.2 effect size is small. ROC is a graphical plot that illustrates the diagnostic ability of a binary classifier system that ranges from 0 to 1. We respectively set hyper-parameters $w_i$, $w_j$, $w_{ij}$ and $T$ as 0.125, 0.125, 0.75, and 8, which were selected by exploratory experiments. As a result, we found that the proposed feature $R$ for people with dementia was significantly higher than that of controls ($p < 1.0 \times 10^{-24}$; Figure 2). We also obtained the effect size of 2.68 (95% confidential interval (CI): 2.11-3.25) and the AUC-ROC of 0.97 (Figure 3).

We next compared $R$ with other features extracted from single conversation that were typically used in previous studies: vocabulary richness, sentence complexity, and repetitiveness. For vocabulary richness, we investigated Type-token ratio, Brunet’s index, and Honoré’s statistic. For sentence complexity and repetitiveness, we respectively employed mean sentence length and sentence similarity. We computed the sentence similarity using cosine distance of sentences defined as TF-IDF (term frequency-inverse document frequency) vectors. Among the six features, we observed significant differences between control and dementia in three features: $R$, Honoré’s statistic, and sentence similarity ($p < 1.0 \times 10^{-24}$, $p < 5.0 \times 10^{-20}$, $p < 5.0 \times 10^{-6}$, respectively; Figure 2). In contrast, mean sentence length, type-token ratio and Brunet’s index had no significant difference between the groups ($p > 0.05$). The proposed feature $R$ showed the best results in terms of effect size and ROC ($d=2.68$, ROC=0.97), followed by Honoré’s Statistic ($d=-1.36$, ROC=0.86), and sentence similarity ($d=0.69$, ROC=0.72) (Figure 3).

Our proposed feature $R$ aims to characterize word repetition, especially in conversations on different days in addition to a single conversation on a single day. We investigated the usefulness of the feature representing the repetition in
Figure 2: Feature distributions for control and dementia in our proposed feature $R$ and the existing five features used in previous studies. Boxes denote the 25th (Q1) and 75th (Q3) percentiles. The line within the box denotes the 50th percentile, while whiskers denote the upper and lower adjacent values that are the most extreme values within Q3+1.5(Q3-Q1) and Q1-1.5(Q3-Q1), respectively. Filled circles show outliers, and squares represent mean values.

conversation on different days. We compared it with the feature extracted from single conversation. Specifically, the former feature was extracted from paired conversational data (i.e. HS$^{-1}$ extracted from $D_{ij}$), while the latter was extracted from single conversational data (i.e HS$^{-1}$ extracted from $D_i$ ($D_j$)). The features extracted from paired conversations and single conversation both showed significant differences between control and dementia ($p < 1.0 \times 10^{-24}$ for single conversation; $p < 1.0 \times 10^{-24}$ for paired conversation). We also found that the features extracted from paired conversations had larger discriminative power than those extracted from single conversations in terms of effect size and ROC score ($d=1.58$, ROC = 0.86 for single conversation and $d=2.67$, ROC = 0.96 for paired conversation; Figure 4). The results suggest that the feature part extracted from paired conversational data contributes to detection performance of our proposed feature.

As an additional analysis, we investigated the tendency of the proposed feature with different intervening days of paired conversations. Specifically, we compared the discriminative power of the feature extracted from paired conversation by changing $T$ from three to 14 days. To define $R$ for paired conversation, we used weight parameter $w_{ij}$ as 1 and others as 0, because we wanted to investigate the relationship between duration $T$ of the paired conversation and word repetitions. In all $T$ values calculated in this study, the proposed $R$ feature for people with dementia was significantly higher than that of controls ($p < 0.05$; Table 2). The effect size and the AUC values respectively ranged from 1.5 to 2.67 and from 0.87 to 0.96. After they increased in the beginning, they peaked at around $T=8$ (effect size of 2.67, 95%
Figure 3: Comparison of our proposed feature $R$ with the existing five features used in previous studies. Error bars are 95% confidence intervals.

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Effect size (95% CI)</th>
<th>ROC</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>R (T=8)</td>
<td>2.68 (2.11, 3.25)</td>
<td>0.97</td>
<td>blue</td>
</tr>
<tr>
<td>Hone's statistic</td>
<td>-1.36 (-1.65, -1.06)</td>
<td>0.86</td>
<td>red</td>
</tr>
<tr>
<td>Sentence similarity</td>
<td>0.69 (0.41, 0.98)</td>
<td>0.72</td>
<td>yellow</td>
</tr>
<tr>
<td>Mean sentence length</td>
<td>0.25 (-0.03, 0.53)</td>
<td>0.64</td>
<td>purple</td>
</tr>
<tr>
<td>Type token ratio</td>
<td>0.05 (-0.23, 0.33)</td>
<td>0.51</td>
<td>green</td>
</tr>
<tr>
<td>Brunet's index</td>
<td>-0.07 (-0.35, 0.20)</td>
<td>0.48</td>
<td>cyan</td>
</tr>
</tbody>
</table>

CI: 2.10-3.24; ROC = 0.96 and had a tendency to decline (Table 2 and Figure 5).

Conculsion

In light of the increasing demand for detecting dementia in everyday situations, we focused on word repetition in separate conversations on different days on the basis of a previous descriptive study and proposed a feature to characterize
To test our proposed feature, we investigated conversational data obtained from a regular monitoring service. The data the service provided was collected from eight elderly people, including two dementia patients, for a period of up to 30 months at the maximum.

First, we found that our proposed feature has strong discriminating power and achieved up to 2.68 for effect size of Cohen’s d and 0.97 for AUC-ROC scores. We also compared our proposed feature with other linguistic features such as vocabulary richness and sentence similarities. As a result, our feature $R$ outperformed other features, suggesting that the use of our feature in addition to already existing feature sets has promise to improve detection performance. We also analyzed feature extracted from single conversational data and paired conversational data. The results indicate that features from paired conversations on different days may be more advantageous in increasing discriminative power than extracting them from a single conversation. In addition, the results obtained in changing the intervening days of two separate conversational data suggest that our feature $R$ could be especially useful when extracted from conversations of the specific intervening days.

Our study has several limitations. One of the limitations of this study is the small sample of participants. Another limitation is its specific type of conversational data. We only analyzed the conversations from a regular monitoring service, where all conversations were over the telephone. We need to investigate our proposed feature with conversational data collected from everyday life situations such as face-to-face family conversations. In addition, our feature treated a binary state of dementia. The main reason is that the sample size of participants was too small to classify dementia in terms of its severity. In future work we will need to collect data from a larger number of participants, which will allow us to test whether or not our feature could be extended to score dementia on a scalar or ordinal scale. However, because to the best of our knowledge this is the first study which aims to characterize word repetitions in conversations on different days, we believe that using our feature will help make it possible to provide service for detecting dementia in daily monitoring.

Acknowledgements

We sincerely thank A. Kamiyama, Y. Masuda, J. Hayakawa, and K. Cho at Cocolomi Co., Ltd. for providing all of the data used in this study, and appreciate valuable comments and suggestions.
Table 2: Discriminative power between control and dementia for our proposed feature $R$ with different $T$.

<table>
<thead>
<tr>
<th>T [days]</th>
<th>Control</th>
<th>Dementia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD)</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>5.31 (0.64)</td>
<td>389</td>
</tr>
<tr>
<td>4</td>
<td>5.35 (0.61)</td>
<td>506</td>
</tr>
<tr>
<td>5</td>
<td>5.34 (0.48)</td>
<td>679</td>
</tr>
<tr>
<td>6</td>
<td>5.38 (0.49)</td>
<td>543</td>
</tr>
<tr>
<td>7</td>
<td>5.48 (0.50)</td>
<td>350</td>
</tr>
<tr>
<td>8</td>
<td>5.52 (0.50)</td>
<td>299</td>
</tr>
<tr>
<td>9</td>
<td>5.49 (0.51)</td>
<td>238</td>
</tr>
<tr>
<td>10</td>
<td>5.44 (0.51)</td>
<td>134</td>
</tr>
<tr>
<td>11</td>
<td>5.39 (0.47)</td>
<td>76</td>
</tr>
<tr>
<td>12</td>
<td>5.30 (0.44)</td>
<td>55</td>
</tr>
<tr>
<td>13</td>
<td>5.40 (0.46)</td>
<td>42</td>
</tr>
<tr>
<td>14</td>
<td>5.45 (0.45)</td>
<td>34</td>
</tr>
</tbody>
</table>

References


Applying Probabilistic Decision Models to Clinical Trial Design

Wade P Smith, Ph.D.1, Mark H Phillips, Ph.D.1,2
1Department of Radiation Oncology, University of Washington, Seattle, WA; 2Biomedical Informatics and Medical Education, University of Washington, Seattle, WA

Abstract

Clinical trial design most often focuses on a single or several related outcomes with corresponding calculations of statistical power. We consider a clinical trial to be a decision problem, often with competing outcomes. Using a current controversy in the treatment of HPV-positive head and neck cancer, we apply several different probabilistic methods to help define the range of outcomes given different possible trial designs. Our model incorporates the uncertainties in the disease process and treatment response and the inhomogeneities in the patient population. Instead of expected utility, we have used a Markov model to calculate quality adjusted life expectancy as a maximization objective. Monte Carlo simulations over realistic ranges of parameters are used to explore different trial scenarios given the possible ranges of parameters. This modeling approach can be used to better inform the initial trial design so that it will more likely achieve clinical relevance.

Introduction

The design of a clinical trial can have a large impact on the progress of medical practice. A number of competing factors need to be considered, such as cohort size, ability to recruit subjects, overall cost, expected size of effect and expected impact of the outcome of the trial. While these and other factors are considered in the design of any trial, there is continuing controversy over how best to do it. One persistent question is whether a clinical trial is solely for the purpose of inference or whether it is a decision problem (for a concise synopsis see Spiegelhalter et al1).

Traditionally, clinical trial designs are a combination of relatively straigh-forward statistics and clinical judgment. The latter informs the design by providing an estimate of the effect size that will alter clinical practice, i.e. there is no point in conducting a trial whose positive outcome will not change how physicians practice. The factors to be considered typically include things like possible increase in side-effects, ease of administering the treatment and cost. If it seems reasonable that the effect to be measured will achieve such a magnitude, then the number of trial subjects is calculated using classical statistics to minimize false positive and false negative trial results. In this approach to trial design, one seeks to infer the true value of the treatment difference between the two arms from the sample obtained through the clinical trial.

When a trial is viewed as an inference problem, the focus is on determining whether the statistics of the trial provide enough certainty to achieve a significant p-value with one, or possibly several related, outcomes. For example, in cancer therapy the common outcomes studied are overall survival, disease-free survival, and local recurrence2. Classic trial design then continues by determining the likelihood function for the chosen outcome, applying the parameters of the control group, agreeing on a clinically relevant difference between new and old treatments, and finally, calculating numbers of subjects given acceptable thresholds of Type I and II errors. Recently, Bayesian inference in clinical trial design has become more common and it can be used for two different purposes. First, using Bayes’ theorem allows for the inspection of the most probable outcomes after only some fraction of the expected number of patients are accrued. This can help to direct resources more efficiently later in the trial and/or to insure that any potential harm is minimized. The other use of Bayesian statistics is to quantify the magnitude of the treatment effect rather than just testing the null hypothesis.

As mentioned, trial design needs to take into account aspects of the new procedure or treatment beyond the main outcome of interest. For example, a clinical trial of a new cancer therapy will be designed to measure the difference in survival at a given time point. However, the future use of the new therapy will also depend on whether more or different complications result, how difficult it is for the patient and/or provider to administer the treatment, and perhaps how much the new treatment costs. These conflicting considerations characterize problems that are best approached using decision theory.

If trial design is viewed as a formal decision problem, then a loss (or utility or cost) function is needed to quantify
the effects, and the expectation value of the utility (or cost) is used in the decision making process. This function is designed to capture the many different aspects of the outcome—the response of the disease, the induced health state, and the feasibility and cost of the new procedure or treatment. The fact that the uncertainties inherent in the problem result in the use of expectation values of the utility indicates that a model of the probabilities of the outcomes is needed. While this adds somewhat to the complexity of framing and quantifying the medical environment compared to the inference problem, it has the benefit of being able to incorporate multiple, competing outcomes. Thus, one obtains a much more complete and clinically realistic view with the aim of designing a trial that will more likely have an impact in clinical decision making.

In this paper, we take the more comprehensive view and model clinical trial design as a decision problem. We combine several probabilistic modeling approaches in order to capture the uncertainties in the disease and treatment response processes and the inhomogeneities within the subject cohorts. The separation of these two sources of variability is an explicit part of our model which is needed to account for their potential effects. In addition, it allows us to generalize our model that was designed for clinical trial design and apply it to clinical decision making, as well. Our model also makes use of a more general utility function, namely quality adjusted life years (QALY). This function captures the interplay between cure and complication as well as integrating the outcomes over time. Overall, our approach offers several advantages over current methods. It allows us to simulate different measured outcomes, thereby giving quantitative guidance as to whether a given trial design is likely to result in clinically useful information. In addition, by calculating the expected value of information regarding different variables, the trial can be designed to either control for certain variables or, prior to initiating the trial, more information can be obtained prior to the trial. Finally, it should be mentioned that our method focuses on the effect size. Traditional trials follow the “null hypothesis significance testing” approach which fails to give any information about the effect size. As our analysis shows, realistic modeling of effect size is a key element to a successful trial.

**Methods**

We have applied our method to clinical trials for HPV-positive (HPV+) oropharyngeal cancer. Subgroup analysis of large clinical trials show that HPV+ patients have significantly improved survival compared to HPV-negative (HPV-) patients under standard radiation therapy protocols. A number of non-inferiority clinical trials have been initiated to determine if lower doses of radiation can be equally efficacious, the goal being to decrease the incidence of severe complications due to the radiation therapy. Thus, the trial design must include a measure of the difference in survival at some relevant time and the differences in complications. In this case, the new treatment uses fewer treatments which is acknowledged to be a benefit to the patient and to incur less cost, so there is no need to model any tradeoff in that area.

**Model framework**

The structure of the model was obtained through in depth discussions with experts in the field of head and neck cancer radiation therapy and from previous publications. We constructed an influence diagram (Figure 1) that incorporates the relevant variables with respect to tumor control and complications to nearby critical structures, along with the decision variables (radiation dose and technique). An influence diagram is a Bayesian network (BN) that includes decision nodes and a utility node. In Figure 1, rectangles denote decision (also known as action) nodes; ovals represent chance nodes; and diamonds represent utility nodes. Color coding of the chance nodes indicate the details of the implementation of joint probability tables. Green nodes indicate prior distributions; red nodes indicate model-generated probabilities; and yellow indicate normal chance nodes.

In a conventional influence diagram, the expected utility (denoted by a diamond-shaped node) is calculated by means of the relevant probabilities obtained from the BN and a utility function. We have expanded the notion of expected utility from the standard form. Instead of using a single function to map preferences and probabilities to an expected utility, we use a Markov model to calculate quality adjusted life years (QALY) (Figure 2). The BN provides the distribution of initial states to the Markov model. The Markov model contains states for status-post RT, recurrent disease, Gr2+ xerostomia, Gr2+ dysphagia, both toxicities, and death. To each state is assigned a utility, and QALYs are accumulated as the Markov model is evaluated annually using a cohort simulation.
Figure 1: Influence diagram modeling the treatment and response of HPV+ oropharyngeal cancer.

In a conventional influence diagram, the optimal action can be solved in a direct fashion. However, the use of the Markov model complicates this process; therefore, we have used a Monte Carlo process to provide values of the probabilities from the BN and to evaluate the Markov model and calculate QALY. These values are then used to choose the optimal action. This approach also allows us to perform a sensitivity analysis as part of the QALY calculation in a natural way. For any given distribution of a variable, the Monte Carlo process samples the distribution that describes the variabilities inherent in its value.

Tumor Control Probability

Modeling the tumor control probability (TCP) is one the most crucial steps. The non-inferiority trials being planned are based on the results reported by Ang et al for HPV+ patients, and our model relies on the same. They stratified patients based on various risk factors, such as smoking and clinical pathology variables (see Figure 1). Most HPV+ patients were in the low risk category (smoking history being a deciding factor), and our current model focuses solely on that group. Ang et al reported a 3 year survival rate of 82.4% [95% CI, 77.2 to 87.6]. Given this variance, a clinically meaningful difference in survival would need to be 5% or greater. The dose reduction trials are looking at the difference in survival between a tumor dose of 70 Gy and a dose of 55 Gy. We modeled three different scenarios, namely that a dose reduction from 70 Gy to 55 Gy would yield either a 7%, 10% or 15% reduction in survival.

TCP is nearly always modeled as a logistic curve with two free parameters, the dose at which TCP = 0.5 [D_{50}] and the slope at that point, \( \gamma_{50} \). We modeled these two parameters as beta functions from which the Monte Carlo process sampled. For each of the three scenarios, the TCP was calculated using the sampled values and only those values were used for which TCP(70 Gy) = 82% and the difference in survival at TCP(55 Gy) \( \in [7, 10, 15]\% \). The means and probabilistic sensitivity analysis ranges (95%) were 34.7, 22.5 – 45.7 Gy and 0.55, 0.24 – 0.98 for \( D_{50} \) and \( \gamma_{50} \), respectively.

Normal Tissue Complication Probability: xerostomia

Normal tissue complication probabilities (NTCP) are calculated by means of logistic functions that are fitted to the data for given grades (or range of grades) of complications to a given organ. Xerostomia (“dry mouth”) occurs when the parotid glands suffer radiation injury. Such complications are common in oropharyngeal cancer with 40-60% of patients suffering significant sequelae. In contrast to the tumor in which the planned dose is uniformly distributed throughout the volume, normal tissues receive a distribution of doses. The actual distribution is most strongly influ-
enced by a particular patient’s anatomy, that is, the geometric relationship between the tumor volume and the organ at risk. NTCP models usually use some summary value, e.g. mean dose, although more complex, radiobiologically-based models are also used\textsuperscript{12}.

To calculate the probability of xerostomia, we used a model published by Miah \textit{et al}\textsuperscript{13}, which uses the mean dose to the parotid gland. This model used non-linear logistic regression analysis to determine the probability of xerostomia from a model with two parameters, D50, the mean dose which induces grade \(\geq 2\) xerostomia in 50\% of patients, and \(k\), the additional rate of xerostomia for each additional Gy at D50. The mean and range on these parameters for the Monte Carlo sensitivity analysis is 42.9 Gy [29-56 Gy] and 1.7 [0.5-2.8]. We also used the Monte Carlo process to sample a distribution of mean parotid doses that were obtained from the treatment planning database of the Department of Radiation Oncology, University of Washington. The mean and range of the parotid mean dose was 21.4 [6-38.9] Gy.

**Normal Tissue Complication Probability: dysphagia**

In our influence diagram probabilities for dysphagia come from an NTCP model developed from a prospective study\textsuperscript{14}. Their work finds values for the logistic function regression coefficients of 0.04 \(\pm\) 0.02 and 0.06 \(\pm\) 0.02 for the SGL and SPCM, respectively; these are used as distributions in the Monte Carlo sensitivity analysis. Their model predicts swallowing dysfunction as a function of mean doses to the supraglottic larynx (SGL) and superior pharyngeal constrictor muscle (SPCM). Possible ranges for mean doses to these structures are dependent on the radiation planning technique employed which varies across clinical practices. We used a range of mean doses appropriate for an IMRT planning technique which attempts to spare the SGL and SPCM\textsuperscript{15}, as this technique is likely to be employed by physicians considering prescription dose de-escalation in order to reduce toxicity. The means and ranges were 52.5 Gy [29.6 – 67.4 Gy] and 63.0 Gy [34.5 – 69.8 Gy] for the SGL mean dose and SPCM mean dose, respectively.

**Utility values**

Multiattribute utilities for complex health states are not easy to obtain and can depend on the method of elicitation. Meregaglia \textit{et al} reviewed a wide range of publications in the area of utilities for health states relevant to head and neck cancers\textsuperscript{16}. We used the values collected by Ramaekers \textit{et al} for our model\textsuperscript{17}. In our model, we sampled from distributions that we feel mainly represent heterogeneity between patients, though an interpretation that the distributions actually represent uncertainty regarding methods is not unassailable.

**Monte Carlo simulations**

Probabilistic sensitivity analyses were conducted using a Monte Carlo sampling technique with 50,000 patients\textsuperscript{10}. By this we mean that a single patient was simulated by selecting at random from each of the distributions in the diagram. The probability of selection of any given value is determined by the functional form of distributions that we used. For example, a normally distributed variable with a given mean and standard deviation would be sampled such that by the end of the analysis, the distribution of values used would match the initial normal distribution. Since each variable is sampled independently, one ends up with 50,000 “patients” whose characteristics are distributed as expected in a heterogeneous population. For each run of 50,000 samples, the QALYs for each patient were calculated and histograms of the results plotted. Separate runs were calculated for each of the different scenarios that represent possible trial outcomes (current knowledge, survival reductions of 7\%, 10\% and 15\%) and for each of the four actions (delivery of 70, 65, 60 and 55 Gy).

**Results**

The results of our simulation using this model are in the form of: (a) a table of expected value of perfect and partial perfect information, (b) histograms of the QALY’s achieved under the different possible trial outcomes and for different delivered tumor doses, and (c) a plot of the fraction of patients who would benefit from each of the possible decisions (tumor doses) as a function of true survival differences.

Table 1 contains the values of expected value of perfect information (EVPI) and expected value of partial perfect
information (EVPPI). The column EVPI presents the results when all variables are assumed known, and places an upper bound on what can be achieved by acquiring perfect information on only a subset of variables (EVPPI). The five columns under the heading EVPPI present the results of partial perfect information, i.e. each variable listed represents the QALYs gained if the given variable is known exactly.

Table 1: Value of information calculations for expected value of perfect information (EVPI) and partial perfect information (EVPPI) (columns 3–7) for different variables. VOI is calculated for four different scenarios: only using our current knowledge, a reduction in cure rate of 15% under a dose reduction of 77 to 55 Gy, a reduction in cure rate of 10% and a reduction in cure rate of 7%.

<table>
<thead>
<tr>
<th>scenarios</th>
<th>EVPI</th>
<th>utilities</th>
<th>TCP</th>
<th>NTCP(xerostomia)</th>
<th>NTCP(dysphagia)</th>
<th>Organ-at-risk doses</th>
</tr>
</thead>
<tbody>
<tr>
<td>current knowledge</td>
<td>617</td>
<td>112</td>
<td>66</td>
<td>0</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>15%</td>
<td>714</td>
<td>239</td>
<td>70</td>
<td>0</td>
<td>94</td>
<td>0</td>
</tr>
<tr>
<td>10%</td>
<td>1175</td>
<td>564</td>
<td>172</td>
<td>24</td>
<td>437</td>
<td>5</td>
</tr>
<tr>
<td>7%</td>
<td>1239</td>
<td>653</td>
<td>31</td>
<td>37</td>
<td>614</td>
<td>22</td>
</tr>
</tbody>
</table>

Figure 2 shows the distributions of QALYs for a range of doses between 70 and 55 Gy. We included 65 and 60 Gy in order to illustrate that this method could be used to guide selection of which doses are chosen as experimental arms. Here the possible trial result that dose de-escalation to 55 Gy results in a less than 7% reduction in overall survival has been entered into the model in order to evaluate which uncertainties contribute the most to the decision uncertainty for a particular patient.

Note that all four panes of Figure 2 are calculated on the same modeled cohort, but with different prescription doses. For some individuals, the expected benefit increases, and for some it decreases, depending on the individual’s probabilities of developing side effects, their preference for the various health states, and their expected probability of survival. The standard deviations of the distributions are reduced as the prescription dose is lowered. In order to illustrate why this occurs we performed a subset analysis on the patients whose QALYs for 70 Gy are greater than the maximum (16.7 QALYs) expected benefit calculated for the 55 Gy arm. None of the individuals in this subset can benefit from a dose reduction to 55 Gy. In this group the average utilities for both side effects were higher than the cohort average, and the average doses to all critical structures were lower than the cohort average, thereby lowering the probability of developing side effects. Patients at these extreme values, are expected to derive the most benefit from maximizing life expectancy, which occurs at the highest dose. Similarly, the group of modeled patients whose expected QALYs for 70 Gy were less than the minimum expected for 55 Gy would benefit from a dose reduction. Patients in this modeled cohort assign lower utilities to health states with toxicity, and also had a higher probability than the cohort average for developing a side effect. This group benefits the most from avoiding toxicity, which occurs at the lowest dose.

Figure 3 shows the distributions of QALYs for the current state of knowledge and for range of possible trial results for the de-escalated dose to 55 Gy. As the difference in overall survival between 70 Gy and 55 Gy decreases, the average population QALYs increases due to the higher survival rate on the 55 Gy arm. The probability of toxicity events is the same for these 4 scenarios, since all are calculated for 55 Gy, and the toxicity and tumor control probabilities are independent probabilities.

The results of the Monte Carlo simulation allow for the calculation of the proportion of patients who would achieve the highest number of QALYs for a given strategy (tumor dose). Figure 4 plots the proportions of patients who benefit from a given dose as a function of the true difference in survival at doses of 70 and 55 Gy, i.e. TCP(70) - TCP(50). At large differences, the decrease in complications cannot overcome the relative importance improved survival. The ordering of the curves at high doses reflect the nature of the logistic TCP curves. At small differences, the advantage of complication sparing is more important as the survival values are closer in value.

For smaller reductions in overall survival (the left side of Fig. 4 or the final row of Table 1) the EVPI increases since decreased doses offer more QALYs to a greater proportion of the population, and perfect knowledge of other model
Figure 2: Histograms of QALY’s calculated under the hypothesis that a reduction in tumor dose from 70 to 55 Gy would result in a 7% reduction in survival. Calculations were performed for four different values of the delivered tumor dose: 70, 65, 60, 55 Gy.

parameters would allow the optimal choice for each individual. With the current state of knowledge, acquiring additional information about any one set of parameters does not have as significant an impact on the decision made, since the remaining variability has a large effect. However, for trial results where dose reduction results in smaller reduction in survival, then acquiring additional information about patient preferences or dysphagia NTCP model parameters has greater potential to change the decision.

Discussion

We have presented a model that is based on two major premises. First, the design of clinical trials should be considered as a decision problem, not merely as an inference problem. Second, the utility function needs to include multiple attributes as well as integration over time. The former was addressed by means of a probabilistic knowledge representation in the form of an influence diagram. The latter was accounted for by using a Markov model to calculate quality adjusted life years. This decision rendered the normal approach to solving the influence diagram for the optimal decision difficult to achieve; therefore, we used a Monte Carlo sampling method.

We applied this clinical trial design model to the case of non-inferiority trials for HPV+ oropharyngeal cancer. Radiation therapy (with and without chemotherapy) of oropharyngeal cancer leads to significant sequelae in a large fraction of patients. The most recent large clinical trial identified patients who were HPV+ as having better survival probability than HPV- patients\(^4\). Given the complication profile, the question arises as whether a reduced radiation dose for HPV+ patients might yield similar survival benefits with reduced complications when compared to current dose protocols. With this in mind, a number of clinical dose de-escalation trials have begun\(^5\).

Classic radiation biology theory holds that a reduction of dose on the order of over 20% will result in a decrease in cell killing, with a concomitant decrease in tumor-related outcomes. What is unknown is whether the dose-response
Figure 3: Histograms of QALY’s achieved if the delivered tumor dose is 55 Gy under four different possible trial outcomes: (a) current protocol with no additional knowledge, (b) 55 Gy results in a 15% reduction in TCP, (c) 55 Gy results in a 10% reduction, and (d) 55 Gy results in a 7% reduction.

The curve is so flat that the outcome is essentially the same. There are two ways that this could come about. Since our data are statistical in nature, non-inferiority can imply confirmation of the standard null hypothesis between the current and the lower dose protocols. The second way that equivalence could be established is to acknowledge that clinical decisions incorporate consideration of both cure and complication, i.e. viewing the trial as a decision problem.

Consider the first possibility, namely that a clinical trial need only confirm the null hypothesis. Ang et al report a standard deviation of 2.5%⁴. Since we are looking at the difference in survival, the ability to distinguish two different mean survival values depends on the sum of the variances. Therefore, in this case, a trial result that could reliably reject the null hypothesis at a level of 0.05 would imply a difference in means of approximately 7%. This consideration led us to our first scenario in which a trial based on inference is on the threshold of significance, that is, able to achieve a p-value of 0.05 or better. Our other scenarios (differences of 10% and 15%) were designed to explore the possibility that reduction in complications could compensate for a larger decrease in survival.

The histograms (Figures 2 and 3) show the changes in the distributions of QALY’s obtained for different segments of the population. They quite nicely illustrate the effects of the inhomogeneities in patient response and geometry. The results for the four different tumor doses show relatively smooth evolution, as would be expected given the logistic curve. They also can help determine the likely outcomes for different true differences in survival.

The gains in QALY as a function of reduction of uncertainty (Table 1) in different variables provides insight into likely sources of variability that could hamper a conclusive trial. The relatively large gains obtained with perfect information about utilities highlights one of the reasons that viewing clinical trials as decision problems is difficult. However, the same table indicates that better knowledge about dysphagia can have a large impact, especially when compared to the other complication, xerostomia.
Figure 4: A plot of the fraction of patients for whom a given tumor dose (70, 65, 60, 55 Gy) would yield the most QALY’s as a function of the difference in TCP(70 Gy) - TCP(55 Gy). The difference is relative to the TCP(70 Gy) reported in Ang et al\textsuperscript{4}.

This approach to modeling clinical trials has weaknesses. When compared to the typical inference approach (described above), our method requires much more quantitative work. To construct an adequate influence diagram, one needs reliable conditional probability tables for all the combinations of parent and child states. While some variables have binary states, some may be continuous variables leading to the imposition of a small number of ranges of values. In all of these cases, it is necessary to use realistic values. In addition, distributions among the various states of variables without parents must be determined. Although this allows an easy method for isolating different cohorts (for example, patients with a given T or nodal stage), one would need to measure, find in the literature, or estimate the distribution that will be present in the trial.

An important distinction between this method and the traditional method is the latter’s reliance on null hypothesis significance testing (NHST). Under this approach, a trial outcome is considered “significant” if it obtains a p-value less than 0.05 (or 0.01 in more rigorous trials). However, there is numerous literature that describes in detail the flaws inherent in relying on such a method\textsuperscript{19,20}. Our method explicitly models the outcomes as a function of effect size, thereby providing a much deeper understanding of what effect size is needed to make a study clinically relevant. This,
in turn, leads to a more accurate calculation of the power of a study. Overall, such an in-depth analysis in trial design will lead to less wasted effort and inconclusive trials.

Overall, modeling the clinical trial as we have provides a number of benefits which extend beyond the clinical trial itself. Such a model provides more insight into the likelihood that a given expected benefit or detriment will actually result in a clinically useful result. It also helps focus on the role that “nuisance” parameters play and the extent to which variability in them can reduce the significance of trial results. Thus, it may be prudent to either wait for better knowledge to be gained about a particular variable or to use that variable to distinguish between different arms of the trial. For example, dose de-escalation trials are more likely to result in significant results if patients whose tumors are near the swallowing muscles are separated from those whose are more distant, particularly if quality of life or dysphagia are to be measured as well.

The use of the simulation process has an added benefit outside of clinical trial design. As Fig. 4 points out, some fraction of patients are more likely to benefit than others from any given decision. Our model allows us to isolate the variables behind such differences, thereby providing clinical decision support.

We note that the distributions that were obtained from this model could be used to help design a clinical trial using Bayesian methods should one so desire. Prior probability distributions can be obtained from the model and incorporated into the design. In this particular case, a Bayesian approach is probably preferable over the standard Neyman-Pearson method. Knowledge of the actual difference in survival benefit, rather than a simple non-inferiority judgement is more likely to help clinicians. This is particularly true given the fact that such trials as these are unlikely to be repeated and reliance on simple p-values may result in unfortunate recommendations given the limited statistical sampling that will occur in a single trial.

**Conclusion**

A model for clinical trial design was described using probabilistic representation of the inhomogeneities within the patient population and the uncertainties regarding treatment response. The clinical trial was viewed as a decision problem and a cost function, in the form of a Markov model calculating quality adjusted life years, was implemented. The model was applied to a current issue in radiation oncology, namely the possible advantage of dose de-escalation in HPV+ oropharyngeal cancer. Using a probabilistic sensitivity approach, results for different possible trial outcomes were calculated in the form of distribution of QALY’s, the value of information, and the fraction of patients who would benefit from different strategies. This approach offers a number of advantages of traditional inference approaches, although it does require more effort up front. The overall goal of this method is to provide information regarding the critical variables so that clinical trials can be designed that will yield the most clinically useful information.

**References**

Inpatient Clinical Order Patterns Machine-Learned From Teaching Versus Attending-Only Medical Services

Jason K. Wang,1 Alejandro Schuler,2 Nigam H. Shah, MBBS, PhD,2 Michael T. M. Baiocchi3, PhD, Jonathan H. Chen, MD, PhD4
1Mathematical & Computational Science Program, Stanford University, Stanford, CA, USA;
2Center for Biomedical Informatics Research, Stanford University, Stanford, CA, USA;
3Prevention Research Center, Stanford University, Stanford, CA, USA;
4Department of Medicine, Stanford University, Stanford, CA, USA

Abstract
Clinical order patterns derived from data-mining electronic health records can be a valuable source of decision support content. However, the quality of crowdsourcing such patterns may be suspect depending on the population learned from. For example, it is unclear whether learning inpatient practice patterns from a university teaching service, characterized by physician-trainee teams with an emphasis on medical education, will be of variable quality versus an attending-only medical service that focuses strictly on clinical care. Machine learning clinical order patterns by association rule episode mining from teaching versus attending-only inpatient medical services illustrated some practice variability, but converged towards similar top results in either case. We further validated the automatically generated content by confirming alignment with external reference standards extracted from clinical practice guidelines.

Introduction
Healthcare in the US often falls short of optimal, evidence-based care, with overall compliance with evidence-based guidelines ranging from 20-80%.1 Even with recent reforms, evidence-based medicine from randomized controlled trials cannot keep pace with the ever-growing breadth of clinical questions, with only 11% of guideline recommendations supported by high-quality evidence.2 This variability and uncertainty in medical practice is further exacerbated by a medical knowledge base that is perpetually expanding beyond the cognitive capacity of any individual.3 A clinician is thus left to synthesize vast streams of information in an attempt to make the best decision for each individual patient. As such, medical practice routinely relies on anecdotal experience and individual expert opinion.

To address these issues, healthcare organizations have increasingly adopted clinical decision support (CDS) systems. CDS aims to reinforce best-practices by distributing knowledge-based content through order sets, templates, alerts, and prognosis scoring systems.4 Here we focus specifically on clinical orders (e.g. lab tests, medications, imaging exams) as concrete manifestations of point-of-care decision making. Computerized provider order entry (CPOE) typically occurs on an “a la carte” basis, where clinicians search for and select orders to trigger subsequent clinical actions (e.g. pharmacy dispensing and nurse administration of a medication, laboratory analysis of blood tests, consultation to a specialist). Because clinician memory and intuition can be error-prone, health system committees manually curate template order sets to distribute standard practice guidelines specific to a diagnosis or medical procedure.5 This top-down approach enables clinicians to draw clinical orders from pre-constructed, human-authored templates when treating common scenarios (e.g. pneumonia, stroke).

Existing approaches to CDS increase consistency and compliance with best practices,6–7 but production of this content is limited in scale by a committee-driven, manual production process that oftentimes requires the collaboration of a multi-disciplinary team of physicians, nurses, and department heads. Once an order set is published and made available to clinicians, ongoing maintenance is required to keep it up to date with new evidence, technology, epidemiology, and culture.8 As such, one of the “grand challenges” in CDS is to automatically generate content by data-mining clinical data sources from the bottom-up.9 In the era of electronic health record (EHR) data, there is an opportunity to create a data-driven CDS system that leverages the aggregate expertise of many healthcare
providers and automatically adapts to the ongoing stream of practice data.10 This would fulfill the vision of a health system that continuously learns from real-world data and translates them into usable, point-of-care information for clinicians. Prior research into data-mining for decision support content includes association rules, Bayesian networks, and unsupervised clustering of clinical orders and diagnoses.11-18 In our prior work, inspired by similar information retrieval problems, we developed a data-driven clinical order recommender engine19 analogous to Netflix and Amazon.com’s “customers who bought A also bought B” system.20 Our engine dynamically generates order recommendations based on real-world clinical practice patterns represented in EHR data.

In psychology, the “wisdom of the crowd” phenomenon purports that the collective assessment of a group of non-expert individuals is generally as good as, if not better than, that of individual experts.21 In the context of data-driven CDS for medical decision making, this translates to training machine-learning models on all available data, including patterns generated by clinicians of all levels of experience, rather than just a subset of data generated by the most experienced few. However, effective medical decision making and patient outcomes could conceivably be compromised if patterns are learned from less experienced fellows, residents, and medical students administering care. In fact, patient outcomes tend to worsen during end-of-year changeover when new, less experienced trainees enter the hospital.22 This phenomenon gives rise to the concern over learning indiscriminately from the “wisdom of the crowd” when the crowd consists of both experienced attending physicians and teaching services that include trainees. Prior studies have examined the outcomes of specific procedures when trainees of differing levels of experience are involved in the patient care process, assessing metrics including readmission rates, mortality, procedural time, etc.23-24 As we seek to evaluate emerging data-driven CDS systems, an essential step is understanding the implications of learning from clinical practices of varying experience and potentially of variable expertise and quality.

In this study, we investigate how clinical orders learned from two distinct clinical settings, a university teaching versus an attending-only service, influence the alignment of clinical order patterns learned from data-mining EHR data to clinical practice guidelines.

Methods

We extracted deidentified, structured patient data from the (Epic) electronic medical record for inpatient hospitalizations in 2013 at Stanford University Medical Center via the STRIDE clinical data warehouse.25 The dataset covers patient encounters from their initial (emergency room) presentation until hospital discharge, comprising >20,000 patients and >6.7 million instances of >23,000 distinct clinical data items.

Data Preparation

The large majority of clinical items are medications, laboratory tests, imaging tests, and nursing orders, while non-order items include lab results, problem list entries, admission diagnosis ICD9 codes, and patient demographics on age, gender, and date of death. Medication data was normalized with RxNorm mappings26 down to active ingredients and routes of administration. Numerical lab results were binned into categories based on “abnormal” flags established by the clinical laboratory, or being outside two standard deviations from the population mean. ICD9 codes were aggregated up to the three digit hierarchy to compress the sparsity of diagnosis categories; original detailed codes were retained if the diagnosis was sufficiently prevalent to be useful. Income levels were inferred from 2013 US census data by cross-referencing each patient’s zip code with the median household income in that region. These pre-processing steps enable us to model each patient as a timeline of clinical item event instances, with each instance mapping a clinical item to a patient at a discrete time point.

With the clinical item instances following the 80/20 rule of a power law distribution, the majority of item types may be ignored with minimal information loss.27 In this case, ignoring rare clinical items with <64 occurrences (~0.004% all instances) reduces the distinct item count to ~12% of the original, while still capturing ~95% of the individual item instances. Removing rare items further avoids spurious results with insufficient data for reliable statistical inference. Common process orders (e.g., vital signs, notify MD, regular diet, transport patient, as well as
most nursing and all PRN medications) were excluded, leaving just over 1,400 candidate clinical orders for consideration.

**Cohort Selection**

Using patient provider information, we prepared two patient cohorts seen by distinct general medicine services: 1) a private attending-only service (n=1774) and 2) a university teaching service (n=3404). University teaching services emphasize both medical education and patient care, thus allowing trainees (e.g. medical students, residents, and fellows) to play an active role in deciding on and administering care plans alongside attendings, nurses, and physician assistants. Although trainees discuss high-level care plans with supervising physicians at least once a day, they are almost exclusively responsible for inputting clinical orders into the EHR system. In contrast, the attending-only service is run exclusively by board-certified physicians.

**Propensity Score Matching**

The private attending-only and university teaching services are within the same hospital system with access to the same pool of resources. However, the private attending service exclusively accepts patients who see clinicians in a local private health plan, while the teaching service admits patients of all health plans, including a small number of uninsured. To minimize potential biases arising from this difference, we conducted propensity score matching to balance the two patient cohorts. Using demographic data (age, gender, ethnicity, income level), initial vital signs recorded before the onset of care (temperature, pulse, respiration), and existing diagnoses as covariates, we applied a logistic regression model to compute the probability \( p \) of each patient’s assignment to the attending-only cohort. The propensity score is then defined as the logit function \( \log \frac{e^x}{1 + e^x} \). Caliper matching on the propensity score resulted in balanced cohorts of 1530 patients each. A caliper threshold of 0.18 was chosen based on \( 0.2 \times \sigma \) where \( \sigma \) = the standard deviation across all propensity scores.\(^{28}\)

**Association Rule Episode Mining**

Using the preprocessed data from patient encounters associated with an input patient cohort, we conducted association rule episode mining for clinical item pairs to capture historical clinician behavior. Our previously described clinical order recommender algorithm\(^{19,29}\) counts co-occurrences for all clinical item pairs occurring within 24 hours to build time-stratified item association matrices. For each pair of items \( A \) and \( B \), the co-occurrence counts accumulated can be represented as \( N_{AB,t} \), the number of patients for whom item \( B \) follows \( A \) within time \( t \), as illustrated by the pseudocode below.

For each patient \( P \):

For each item \( A \) that occurs for patient \( P \) at time \( t_A \):

For each item \( B \) that occurs for patient \( P \) at time \( t_B \) where \( t_B \geq t_A \):

If \( (P,A,t_A) \) or \( (P,B,t_B) \) not previously analyzed:

If \( (t_B-t_A) \leq \) timeThreshold and \( (P,A,B) \) newly encountered:

Increment \( N_{AB,\text{timeThreshold}} \)

Record \( (P,A,B) \) as previously encountered

Record \( (P,A,t_A) \) as previously analyzed

These counts are then used to populate 2x2 contingency tables to compute association statistics such as baseline prevalence, positive predictive value (PPV), relative risk (RR), odds ratio (OR), and \( P \)-value by chi-square test with Yates’ correction for each pair of clinical items. For a given query item (e.g. admission diagnosis), we generate a list of clinical item suggestions score-ranked by a specified association statistic. Score-ranking by PPV prioritizes orders that are likely to occur after the query items, while score-ranking by \( P \)-value for items with odds ratio >1 prioritizes orders that are disproportionately associated with the query items.\(^{29}\)
We trained two distinct association models either using patient encounters from the balanced attending-only or teaching service cohorts. We then generated a predicted order list ranked by PPV from each model for 6 common diagnoses: altered mental status (ICD9: 780), chest pain (ICD9: 786.5), gastrointestinal (GI) hemorrhage (ICD9: 578.9), heart failure (ICD9: 428), pneumonia (ICD9: 486), and syncope and collapse (ICD9: 780.2).

Guideline Reference Standard
To develop an external reference standard for order quality, a board-certified internal medicine physician manually curated reference lists based on clinical practice guidelines available from the National Guideline Clearinghouse (www.guideline.gov) and PubMed that inform the inpatient management of altered mental status,30 chest pain,31-32 GI hemorrhage,33-35 heart failure,36-37 pneumonia,38-39 and syncope and collapse.40 The specific diagnoses were selected based on the existence of relevant guidelines and a significant quantity of clinical data examples. Candidate clinical orders were included in the reference standard based on whether a guideline text explicitly mentioned them as appropriate to consider (e.g., treating pneumonia with levofoxacin), or heavily implied them (e.g., bowel preps and NPO diet orders are implicitly necessary to fulfill explicitly recommended endoscopy procedures for GI bleeds).

Evaluation Metrics
PPV-ranked predicted order lists were generated from both attending-only and teaching service association models. To assess the similarity between the two predicted order lists, traditional measures of list agreement like Kendall’s τ -metric41 are not ideal as they often require identically sized, finite lists, and weigh all list positions equivalently, neglecting rank-order. To compare ranked clinical order lists, we instead calculate their agreement by Rank Biased Overlap (RBO).42 RBO computes the average fraction of top items in common between two ordered list and is characterized by a “persistence” parameter p, the probability that an observer reviewing the top k items will continue to observe the (k+1)-th items. For our calculations, we used a default implementation parameter p of 0.98. This has the effect of geometrically weighting emphasis to the top of each list. RBO values range from 0.0 (no correlation or random list order) to 1.0 (perfect agreement of list order).

To obtain a more objective measure of quality, we also compare each predicted order list against the corresponding guideline reference list using area under the receiver operating characteristic (ROC AUC = c-statistic) and precision and recall for the top K ranked items. Comparison of such metrics will determine whether attending-only and teaching service cohorts differ in their practice of guideline aligned medicine. Non-meaningful, small absolute differences may still yield “statistically significant” P values given sufficiently large data sizes; thus, we can judge clinical settings as being comparable if their c-statistics are “bioequivalent.” That is, if the 95% confidence interval for each cohort’s c-statistic falls within 80-125% of the other cohort’s c-statistic (definition from pharmacologic bioequivalence of generic versus brand drugs).43

Results
The post-matching standardized mean differences (SMD) and p-values computed using two-sample t-tests were < 0.1 and > 0.15, respectively, across all covariates, demonstrating a statistically insignificant difference between balanced attending-only and teaching service patient cohorts. Figure 1 shows the SMD of all 25 covariates before and after propensity score matching. Post-matching results indicate that matching was necessary and effective in balancing patient cohorts for a fair comparison with respect to measured covariates.

Table 1 illustrates examples of the top clinical orders predicted by our recommender engine for an admission diagnosis of pneumonia and altered mental status trained separately on each patient cohort. The predicted order lists for pneumonia corroborate on several top clinical items including intravenous administration of levofoxacin, a drug commonly used to treat bacterial infections, and blood cultures to detect pathogens in the bloodstream. Likewise, the two predicted order lists for altered mental status agree on conducting a CT scan of the
patient’s head as well as a screening of blood, urine, or other body samples to investigate the patient’s usage of certain drugs. To assess the similarity between order lists outputted from the attending-only and teaching service association models while still capturing the score-based ranking within each list, we use RBO. RBO values of −0.7 for all 6 common diagnosis shown in Table 2 demonstrate strong agreement between each pair of predicted order lists. Figure 2 shows ROC curves generated for the 6 admission diagnoses. Each plot reports 3 order lists compared against the guideline reference standard, corresponding to 3 curves: attending-only predicted orders, teaching service predicted orders, and the real-world pre-authored order set manually curated by Stanford University’s Medical Center serving as a benchmark. Pre-authored order sets have no inherent ranking or scoring system to convey relative importance and are thus depicted as a single discrete point on the ROC curve. Area-under-curve (AUC) is reported as c-statistics with 95% confidence intervals empirically estimated by bootstrap resampling with replacement 1000 times. Figure 3 depicts recommendation accuracy for an increasing number of K items considered, illustrating the tradeoff between precision and recall.

Figure 1. The standardized mean difference (SMD) between attending-only and teaching service cohorts across 25 covariates spanning demographic data, initial vital signs, and existing diagnoses. For a given covariate, the SMD is defined as the difference between the mean value for each cohort divided by the pooled standard deviation.

Table 1. Top five ranked clinical order associations for pneumonia (ICD9: 486) (top) and altered mental status (ICD9: 780) (bottom) predicted by attending-only and teaching service trained models sorted by P-value calculated by Yates’ chi-squared statistic. Additional association statistics (e.g. baseline prevalence, PPV, RR) and a column denoting the presence or absence of the predicted item in the corresponding human-authored hospital order set or guideline reference standard are also included. Items with a baseline prevalence <1% are excluded to avoid statistically spurious results and ensure computationally tractable association rule episode mining. Each item represents a clinical order that a clinician can request through a CPOE system. An automated order set can be curated by selecting the top K ranked clinical orders.

<table>
<thead>
<tr>
<th>Attending-Only Service Orders</th>
<th>Prevalence</th>
<th>PPV</th>
<th>RR</th>
<th>P-Value</th>
<th>Order Set/ Guideline</th>
<th>Teaching Service Orders</th>
<th>Prevalence</th>
<th>PPV</th>
<th>RR</th>
<th>P-Value</th>
<th>Order Set/ Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respiratory Isolation</td>
<td>0.02</td>
<td>0.18</td>
<td>8.4</td>
<td>2x10^-4</td>
<td>No/No</td>
<td>Azithromycin (Intravenous)</td>
<td>0.10</td>
<td>0.41</td>
<td>4.7</td>
<td>4x10^-14</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Levofloxacin (Intravenous)</td>
<td>0.14</td>
<td>0.46</td>
<td>3.7</td>
<td>9x10^-14</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td>Blood Culture (Aerobic &amp; Anaerobic Bottles)</td>
<td>0.50</td>
<td>0.90</td>
<td>1.9</td>
<td>5x10^-7</td>
<td>Yes/Yes</td>
<td>Respiratory Nebulizer</td>
<td>0.26</td>
<td>0.52</td>
<td>2.1</td>
<td>3x10^-6</td>
<td>Yes/No</td>
</tr>
<tr>
<td>Blood Culture (2 Aerobic Bottles)</td>
<td>0.49</td>
<td>0.88</td>
<td>1.8</td>
<td>1x10^-4</td>
<td>Yes/Yes</td>
<td>Blood Culture (Aerobic &amp; Anaerobic Bottles)</td>
<td>0.59</td>
<td>0.87</td>
<td>1.5</td>
<td>4x10^-4</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td>Respiratory Culture</td>
<td>0.11</td>
<td>0.30</td>
<td>2.9</td>
<td>3x10^-4</td>
<td>Yes/Yes</td>
<td>Acetaminophen (Rects)</td>
<td>0.02</td>
<td>0.11</td>
<td>6.3</td>
<td>5x10^-4</td>
<td>Yes/No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attending-Only Service Orders</th>
<th>Prevalence</th>
<th>PPV</th>
<th>RR</th>
<th>P-Value</th>
<th>Order Set/ Guideline</th>
<th>Teaching Service Orders</th>
<th>Prevalence</th>
<th>PPV</th>
<th>RR</th>
<th>P-Value</th>
<th>Order Set/ Guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT Head</td>
<td>0.22</td>
<td>0.62</td>
<td>2.9</td>
<td>2x10^-14</td>
<td>Yes/Yes</td>
<td>CT Head</td>
<td>0.25</td>
<td>0.63</td>
<td>2.7</td>
<td>6x10^-14</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td>Drugs of Abuse Screen Urine</td>
<td>0.06</td>
<td>0.31</td>
<td>3.7</td>
<td>1x10^-7</td>
<td>Yes/Yes</td>
<td>Ammonia Plasma</td>
<td>0.08</td>
<td>0.26</td>
<td>3.9</td>
<td>1x10^-11</td>
<td>Yes/Yes</td>
</tr>
<tr>
<td>Consult to Neurology</td>
<td>0.03</td>
<td>0.15</td>
<td>5.2</td>
<td>8x10^-4</td>
<td>No/Yes</td>
<td>Volatile Screen</td>
<td>0.09</td>
<td>0.22</td>
<td>2.6</td>
<td>1x10^-3</td>
<td>No/Yes</td>
</tr>
<tr>
<td>Acetaminophen Serum</td>
<td>0.03</td>
<td>0.15</td>
<td>5.2</td>
<td>8x10^-4</td>
<td>No/No</td>
<td>Drugs of Abuse Screen Urine</td>
<td>0.17</td>
<td>0.32</td>
<td>2.1</td>
<td>3x10^-4</td>
<td>No/Yes</td>
</tr>
<tr>
<td>Salicylate Level</td>
<td>0.03</td>
<td>0.13</td>
<td>5.8</td>
<td>8x10^-4</td>
<td>No/No</td>
<td>Rifaximin (Oral)</td>
<td>0.03</td>
<td>0.08</td>
<td>3.7</td>
<td>1x10^-3</td>
<td>No/No</td>
</tr>
</tbody>
</table>
Table 2. Rank Biased Overlap (RBO)\(^2\) computed between attending-only and teaching service order lists, score-ranked by PPV, predicted for 6 common diagnoses: altered mental status (ICD9: 780), chest pain (ICD9: 786.5), gastrointestinal (GI) hemorrhage (ICD9: 578.9), heart failure (ICD9: 428), pneumonia (ICD9: 486), and syncope and collapse (ICD9: 780.2). RBO computes the average fraction of top items in common, geometrically weighting all 1468 or 1474 candidate clinical order items based on a scoring metric (e.g. PPV) for the attending-only and teaching service cohorts, respectively. RBO values of ~0.7 indicate strong overlap between order lists generated by the two cohorts.

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>Rank Biased Overlap Attendee-Only vs. Teaching Service</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altered Mental Status (780)</td>
<td>0.74</td>
</tr>
<tr>
<td>Chest Pain (786.5)</td>
<td>0.72</td>
</tr>
<tr>
<td>Gastrointestinal Hemorrhage (578.9)</td>
<td>0.72</td>
</tr>
<tr>
<td>Heart Failure (428)</td>
<td>0.68</td>
</tr>
<tr>
<td>Pneumonia (486)</td>
<td>0.72</td>
</tr>
<tr>
<td>Syncope and Collapse (780.2)</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Figure 2. ROC plots for the 6 common diagnoses. Each plot compares an order set authored by the hospital and automated predictions from attending-only and teaching service association models against the guideline reference standard. In all cases excluding heart failure, both model-predicted order lists show substantially larger c-statistics than the respective order set benchmark. As the manually-curated hospital order set has no inherent ranking, it is plotted as a single point in which all order set items are considered.
Figure 3. Precision (top) and recall (bottom) curves for 3 common diagnoses: pneumonia (ICD9: 486), gastrointestinal hemorrhage (ICD9: 578.9), and chest pain (ICD9: 786.5). Prediction accuracy (precision or recall) for predicting guideline reference orders is shown as a function of the top K recommendations considered (up to 250) using PPV as the scoring metric. Data labels are added for K = 10 and nO = number of items in the respective hospital order set. nO = 52, 43, and 32 for pneumonia, gastrointestinal hemorrhage, and chest pain, respectively. As the manually-curated hospital order set has no inherent ranking, orders are randomly sampled with replacement from the order set as the curve progresses from left to right.

Discussion

In Table 1, we see that the top clinical orders predicted by attending-only and teaching service association models corroborate on several key items (e.g. levofoxacin, blood culture, CT head, etc.). Although the specific rankings and association statistics vary, as a discrete ranked list the order predictions yield comparable c-statistics (AUC) against the guideline reference standard with 95% confidence intervals satisfying the definition of “bioequivalence” across the 6 admission diagnoses (Figure 2). Both association models outperform hospital order sets for 5 of the 6 diagnoses, with P<10^-100 for all differences between predicted and benchmark c-statistics. In the case of heart failure, the teaching service model performed marginally worse. This instability can be attributed to small data size (n=38 for teaching service patients admitted under heart failure in 2013). In comparison, the teaching service cohort has substantially more (n>100) patients admitted for each of the 5 other diagnoses. Notably, in this hospital’s teaching setting, heart failure patients are largely separated to a specialized heart failure teaching service away from the general medicine teaching service we investigate in this study. Thus, the distribution of patients and order patterns fed into the recommender algorithm may only capture an unusually distinct subset of heart failure admissions. In comparing the similarity between order lists predicted by the two patient cohorts, we see strong average overlap as indicated by RBO values of ~0.7 in Table 2. In relation to a prior study,^20 the attending-only and teaching service order lists across all 6 diagnoses share higher RBO values (greater stability and overlap) than learned order lists for 43 common admission diagnoses compared over time (e.g. order lists generated from 2009 versus 2012 EHR data). Precision and recall curves at varying values of K in Figure 3 pay particular attention to the top items that a clinician
user could realistically be expected to review in a CPOE system. In the case of gastrointestinal hemorrhage and chest pain, we see that attending-only and teaching service order lists achieve greater recall and precision particularly at small values of K. These results support the argument that aggregating clinical order patterns from teaching services characterized by physician-trainee teams and attending-only services characterized exclusively by board-certified physicians will converge towards comparable top results as both cohorts share a common end goal of patient care. This study is an important step in addressing the concern of training machine-learning models on the the aggregate wisdom of clinicians, when the clinical setting contains both trainees and more experienced physicians.

Although propensity score matching was conducted to minimize the influence of confounding covariates recorded in EHR data, reducing the SMD between all recorded covariates to <0.1 as shown in Figure 1, learned clinical order patterns can be influenced by undocumented biases, not explicitly accounted for in demographic data, past medical history, or initial vital signs. In particular, existing hospital order sets are provided as a resource for clinicians to utilize. Automated association models may simply recapitulate the pre-authored order set templates in lieu of truly capturing individual clinical practice patterns. However, only 15% of all clinical orders recorded in the STRIDE dataset were input as part of an orderset. Furthermore, in a supplementary sensitivity analysis, we compared association models trained with and without items input as part of hospital order sets for altered mental status, chest pain, and pneumonia. The resultant ROC curves were largely indistinguishable, providing reassurance that the learned practice patterns were not overly sensitive to or undermined by physicians using pre-existing order set templates. Propensity score matching was conducted on the full attending-only and teaching service patient cohorts. However, when given a query diagnosis (e.g. pneumonia), the clinical recommender engine considers only co-occurrence counts and subsequent association statistics between the specific diagnosis and all other candidate clinical items. Thus, even though the cohorts are balanced with respect to all patients, the subset of patients with a specific diagnosis may not be. Propensity score matching between the two cohorts on a per-diagnosis basis would further reduce potential bias.

In our previous work, we already demonstrated that these methods can accurately predict real-world clinical practice patterns. However, a fair concern is whether real-world practices are synonymous with preferred ones. With the variability and uncertainty of medical practice, it can be perilous to define a gold standard to assess medical decision making. An important contribution of this study is that we evaluate our association models against an external reference standard of clinical practice guidelines to more closely align with patient outcomes and clinical trial evidence. The practice pattern learning methods easily extend to more varied clinical scenarios, but would not be appropriate for this study setting as we specifically focused on admission diagnoses with existing clinical practice guidelines and hospital order sets to enable our benchmark evaluations. Even clinical practice guidelines from trusted sources (e.g. National Guideline Clearinghouse) are published in the form of non-prescriptive clinical text that is deliberately open to interpretation. In this study, guideline reference standards were extracted by a single physician. To improve robustness of the curation process, reference standards can be generated by adjudicating lists generated by multiple physicians independently reviewing clinical practice literature. With the emergence of data-driven CDS systems rooted in machine-learning, it will become increasingly important to advance similar learning and evaluation methods as here for curating standard references for decision making quality. Indeed, clinical practice guidelines are imperfect. The fact that a clinical order was not mentioned in a guideline does not indicate it is an incorrect medical decision. However, defining “appropriate” orders for a given admission diagnoses is nearly impossible in the general case as there is no gold standard in medicine. As such, in this study and in future work, we seek to introduce a variety of different perspectives to evaluate machine-generated clinical order suggestions. Here, we ask whether experienced clinicians yield different practice patterns than trainees by comparing practice patterns learned from two distinct clinical settings. In future work, we focus on concrete patient outcomes by evaluating practice patterns learned from clinicians with substantially higher or lower observed versus expected patient mortality rates.
Conclusion
When extracting patterns from the ongoing stream of practice data made available in EHR systems at academic hospitals, we are implicitly accepting the decision making of less experienced trainees alongside more experienced clinicians. Clinical order recommender systems trained on distinct patient cohorts, one seen exclusively by an attending-only service and the other seen by a teaching service characterized by physician-trainee teams, yield comparable top aggregate results that align with clinical practice guidelines as well as if not better than manually curated decision support content.

References
Evaluation of Flowsheet Documentation in the Electronic Health Record for Residence, Living Situation, and Living Conditions

Tamara J. Winden, PhD1,4, Elizabeth S. Chen, PhD5, Karen A. Monsen, PhD, RN3, Yan Wang, PhD1, Genevieve B. Melton, MD, PhD1,2

1Institute for Health Informatics, 2Dept of Surgery, and 3School of Nursing, University of Minnesota, Minneapolis, MN;
4University of Kansas Medical Center, Kansas City, KS;
5Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

Social determinants of health (SDOH) are important considerations in diagnosis, prevention, and health outcomes. However, they are often not well documented in the EHR and found primarily in unstructured or semi-structured text. Building upon previous work, we analyzed all flowsheet data in 2013 for information related to the SDOH topic areas of Residence, Living Situation, and Living Conditions. Overall, 91 rows were identified as being related to the topics areas resulting in 604,616 unique observations. Individual rows contained SDOH data often covered multiple concepts especially free-text entries. These data included most often references to the residence, residence details, and with whom the patient lives. Very few contained living condition references. Additionally, there was significant duplication and inconsistency of row labels, as well as variation in value list content for rows collecting the same concepts. Our findings demonstrate significant opportunities to improve and achieve better standardization in documentation around these SDOH.

Introduction

Social and individual behavioral factors play an important role in diagnosis, prevention, health outcomes, and quality of life.1-4 Social determinants of health (SDOH) in many instances have a profound impact on our overall health. SDOH includes behavioral components, such as alcohol, drug, and tobacco use, as well as physical environmental factors that can influence the patient’s health such as living conditions, social support, occupation, and physical activities.1 They have been shown to contribute to mortality, as well as a causal mechanism for disease.5-9 Insecurity related to housing has been associated with poor health among children,10 barriers related to access to health care,11 and chronic disease management.12 Living situation, such as residence type, with whom the patient lives, housing density, physical living conditions, and social support, all have been shown to have significant impact on a patient’s health outcomes.13, 14 The National Academy of Medicine (NAM), formerly Institute of Medicine (IOM), suggests that a health policy framework implemented around SDOH would achieve better population health, less inequality, and lower costs.15 There are many public health related initiatives aimed at addressing social determinants and health outcomes as well as recommendations by NAM for SDOH documentation in the EHR.16-18

While much work has been done to demonstrate the effect of behaviors such as alcohol and tobacco use on health outcomes,19 the majority of other social determinants have not been investigated as thoroughly. In general, housing insecurity has been associated with poor health in children, has been shown to have an impact on chronic disease management, and is linked to depression, stress, smoking, and drug use.10-12 Housing, for example, has been studied with respect to the impact of homelessness on various conditions and housing related exposures.20-28 However, little work has been done around examining the health affects of housing density or with whom the patient lives.29 Knowledge regarding the patient’s physical living space, with whom the patient lives, and related exposures would benefit clinicians and other stakeholders in providing appropriate care as well as facilitation of appropriate housing interventions when needed.30

The tremendous increase in the use of EHRs provides unprecedented opportunity to collect and analyze patient data. Even though SDOH documentation has increased in the EHR, social history information is often lacking or inadequate despite its fundamental role in understanding the context of the patient’s story.31 Having SDOH data available in a format that can be analyzed and reused for clinical care along with other associated clinical data would enable development of more informative clinical decision support tools, better evaluation of patient outcomes, generation of additional evidence based care guidelines, and identification of patients who may benefit from special services or interventions or those who may be at higher risk for potentially preventable events.
However, well-designed discrete data collection tools for many aspects of social history information have not been widely developed and incorporated to the EHR. Currently, our work has shown that SDOH documentation may be documented as structured data or unstructured text (e.g., in clinical notes or free-text data collection fields). Moreover, end-users and healthcare institution are left to utilize ill-fitting tools for data collection such as flowsheets. Also, it is common practice for health care organizations to customize their EHR, such as work building flowsheets or other tools to fit specific workflows. In some cases, this customization results in SDOH information being documented in various areas of the chart ultimately making this information difficult for providers to locate. Lastly, comprehensive standards have not yet been developed to encompass all the SDOH topic areas especially the topic areas of interest in this work, namely: Residence, Living Situation, and Living Conditions. In previous work, we found overall that Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) was the most comprehensive source covering these three topic areas followed by the Omaha System. However, both of these sources lacked substantial detail pertaining to the three target topic areas of interest in this work. There is ongoing work to map flowsheet rows to coding systems such as SNOMED CT and Logical Observation Identifiers Names and Codes (LOINC®). However, those data models are still under development and, as such, they are not quite ready for widespread use in clinical documentation.

Flowsheets are critical tools in the EHR for documenting longitudinal data and information such as assessments and observations, as well as providing checklists for routine care tasks. Nurses and allied health staff, such as social workers, physical therapists, and occupational therapists are the typical primary documenters and users of flowsheet data. Since they are used by many different members of the interdisciplinary care team, they naturally provide a mechanism for communication between team members as well as the ability to combine the documentation from disparate disciplines into a single format for visualization.

In the Epic EHR, flowsheets are very flexible in design and fairly easy to build and implement; thus, they offer a popular option for building discrete data collection tools. A flowsheet “Row” is equivalent to a field in a database designed to collect “Observations” or data points. Each flowsheet row can be formatted to collect measure values (e.g., as free-text, numeric, or selection from a custom list) and each row also allows for the entry of free-text comments to accompany each measure value. Flowsheet displays are similar to spreadsheets and are intended to document brief results. Longer text strings or larger blocks of free-text cannot be displayed completely, and therefore these data are cumbersome to review in this format. Lastly, information entered as measure value comments are free-text blocks and are not accessible for typical clinical decision support purposes or for secondary use in population health research. Since flowsheets are so flexible and easy to implement, they can be overused or sometimes misused. The sheer number of rows and the organizational structure can get unwieldy unless care is taken in planning and design of these tools to ensure that each row is unique to capture a specific discrete measurement or concept.

This study is focused on the three target topic areas of Residence, Living Situation, and Living Conditions. Using definitions developed from previous work, Residence describes dwelling types, physical residence, and geographic location and includes safety considerations such as railings or number of floors and steps. Living Situation describes with whom the patient lives such as roommates, family members, multi-resident dwelling as well as how many others they live with. Lastly, Living Conditions describes environmental cleanliness and precautions against infection and disease and includes such things as animals, and presence of mold or an unclean living space.

The goals of this descriptive study were to: (1) examine flowsheet rows relevant to the three target topic areas and (2) characterize flowsheet documentation content and compare flowsheet content to that found in previous work, which analyzed existing standards and terminologies as well as unstructured text.

**Methods**

**Data Sources**

The data used in this study originated from the Fairview Health System (FHS) EHR system and included only data from inpatients from 2013 who had consented for their medical records to be used in research. The primary data source for this work was the Academic Health Center Information Exchange (AHC-IE) data repository, which contains clinical data from the University of Minnesota (UMN), Academic Health Center (AHC), FHS, University of Minnesota Physicians (UMP), and other external data sources (e.g., geocoding data and Minnesota Department of Health death data). Our data exploration focused on EHR data for inpatient visits for 2013 for both FHS and UMP.

**Flowsheet Review**
A complete dataset of all unique flowsheet row unique IDs, system names, and display names, were extracted from the AHC-IE including rows that had been retired. System Names are names given to the rows which are not visible to the end user whereas the Display Names are visible to the end user when they are documenting. The dataset was systematically searched using the list of search terms compiled from previous work, which reviewed existing interface terminologies, standards, specifications, coding terminologies, vocabularies, documentation guidelines, measures, and surveys. The complete dataset of flowsheet rows was then also manually reviewed to ensure complete capture of rows that may have been missed with the search term list and related rows identified. The final set of search terms obtained from the flow sheet data review are summarized in Table 1.

### Table 1. Final set of search terms for flow sheet review.

<table>
<thead>
<tr>
<th>Search terms from standards</th>
<th>Added search terms from flowsheet review</th>
</tr>
</thead>
</table>

**Figure 1:** Flowsheet review methods.

For all rows identified as being related to the three topic areas, observation measure values and observation free-text comments were extracted. Each of the flowsheet rows and observations were examined to ensure relevance to one or more of the three topic areas and unrelated rows and observations were excluded. Rows and observations found to not be related to one or more of the three topic areas were removed from the dataset the final set of observation measure values and observation comments were reviewed. Rows were classified into target domain categories, and sorted based on data collection type, i.e., rows designed to be entered using a drop-down value list or by the entry of free-text. For those rows utilizing drop-down value lists, the complete value list values were extracted, combined, and sorted to develop harmonized value lists for the elements. Meta-data for rows utilizing value lists that included frequency of value lists updates were examined. Measure values from rows that were free-text entry as well as comment text were compiled and reviewed. Lastly the overall flowsheet documentation content was compared to content findings from previous standards work (Figure 1).\(^{32,33}\)

**Results**

**Flowsheet Observations**

The 28,621 unique flowsheet rows system and display names were analyzed systematically using search terms (Table 1). The initial analysis yielded an initial set of 567 rows that were potentially relevant to the target topic areas. After manual review of system and display names, rows found to be unrelated to the topic areas were removed.
resulting in a final list of 222 unique flowsheet rows that could potentially contain data related to the three topic areas of interest.

Flowsheet observation data were extracted from the AHC-IE database for these 222 rows resulting in an initial dataset containing 198 unique flowsheet rows with a total of 1,170,722 observations. Twenty-four of the 222 flowsheet rows were not documented on in the study period so, while they were in the list of rows, they were not represented in the final set of observations. This observation dataset represented 169,425 unique patients and 279,028 encounters. The 198 unique flowsheet rows were reviewed for relevance to the three topic areas and 107 rows were removed from the dataset leaving a total of 91 unique flowsheet rows found to be related to one or more of the three target topic areas (Table 2).

Of the 91 unique flowsheet rows, 49 were found to have been built using custom lists, i.e., pre-built value lists that the user can select from. The remaining 42 rows were built as free-text measure values. Of the 49 rows built with custom lists, 41 were designed to allow the user to select multiple values from a custom list. The 49 rows with custom lists accounted for a total of 480,013 observations where there was either a measure value or comment documented. For 50,066 observations, the measure value was null but there was a free-text comment entered. The 42 rows built as free-text fields resulted in 124,603 observations, i.e., observations that had a value, comment, or both. Of these, 18,932 observations had a null measure value but had a free-text comment entered.

Table 2. Summary of Flowsheet Rows, Observations, and Values.

<table>
<thead>
<tr>
<th>Flowsheet Rows (Unique data capture fields)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Unique Flowsheet Rows</td>
<td>28,621</td>
</tr>
<tr>
<td>Total rows identified with search terms and manual review</td>
<td>567</td>
</tr>
<tr>
<td>Total “cleaned” rows identified through search terms</td>
<td>222</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observations (Data points collected for the identified, 222 flowsheet rows)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total observations for identified rows</td>
<td>1,170,722</td>
</tr>
<tr>
<td>Unique rows with observations</td>
<td>198</td>
</tr>
<tr>
<td>Total unique patients</td>
<td>168,425</td>
</tr>
<tr>
<td>Total unique encounters</td>
<td>279,028</td>
</tr>
<tr>
<td>Total observations after manual review</td>
<td>604,616</td>
</tr>
<tr>
<td>Total unique flowsheet rows</td>
<td>91</td>
</tr>
<tr>
<td>Total unique patients</td>
<td>96,383</td>
</tr>
<tr>
<td>Total unique encounters</td>
<td>139,729</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Values (Data types and presence of associated comments)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Rows built with custom lists</td>
<td>49</td>
</tr>
<tr>
<td>Total observations (Value and/or Comment entered)</td>
<td>480,013</td>
</tr>
<tr>
<td>Total rows with Value only</td>
<td>429,947</td>
</tr>
<tr>
<td>Total rows with Comment only</td>
<td>50,066</td>
</tr>
<tr>
<td>Rows built as free-text fields</td>
<td>42</td>
</tr>
<tr>
<td>Total observations (Value and/or Comment entered)</td>
<td>124,603</td>
</tr>
<tr>
<td>Total rows with Value only</td>
<td>105,671</td>
</tr>
<tr>
<td>Total rows with Comment only</td>
<td>18,932</td>
</tr>
</tbody>
</table>

Table 3 characterizes the final dataset comprised of 91 unique flowsheet rows identified after manual review and the total count of observations found for each row during the study period. The Row Format column indicates the data type of the row measure value observation, i.e., whether the row is a free-text entry or uses a value list, i.e. hard coded list of options to select from with single or multiple select. This table is sorted alphabetically by Measure Display Name to show the duplication in concepts being collected via separate flowsheet rows and variation in the way identical measures are being collected either by formatted list of single- and multi-select options or by free-text. Each row measure value was categorized according to topic area as were free-text comments. Many rows covered more than one of the three topic areas as shown in the last column “Total Topic Areas”.

239
<table>
<thead>
<tr>
<th>ID</th>
<th>Measure Display Name</th>
<th>Obs Count</th>
<th>Row Format</th>
<th>Measure Value</th>
<th>Free-text Comments</th>
<th>Primary Topic Areas</th>
<th>Total Topic Areas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RS LS LC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RS LS LC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RS LS LC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Current Living Environment</td>
<td>573</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2</td>
<td>Current/Future Living Arrangements</td>
<td>3</td>
<td>FT</td>
<td>**</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>Living Arrangement Comments</td>
<td>1638</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>4</td>
<td>Living Arrangement Concerns</td>
<td>13899</td>
<td>VLS</td>
<td>**</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>5</td>
<td>Living Arrangement Concerns</td>
<td>1056</td>
<td>VLS</td>
<td>**</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>Living Arrangements</td>
<td>6978</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>Living Arrangements</td>
<td>8569</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>8</td>
<td>Living Arrangements</td>
<td>1165</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>Living Arrangements</td>
<td>61448</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>Living Arrangements Comment</td>
<td>36</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>Living environment</td>
<td>1149</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>12</td>
<td>Living environment</td>
<td>61</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>Living environment</td>
<td>191</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>Living environment</td>
<td>3741</td>
<td>VLS</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>15</td>
<td>Living environment</td>
<td>1</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>16</td>
<td>Living environment</td>
<td>617</td>
<td>VLS</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>Living environment</td>
<td>2</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>18</td>
<td>Living environment</td>
<td>5</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>19</td>
<td>Living environment</td>
<td>282</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td>Living environment</td>
<td>30</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>Living environment</td>
<td>53</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>22</td>
<td>Living environment</td>
<td>1227</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>23</td>
<td>Living environment</td>
<td>281</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>24</td>
<td>Living environment</td>
<td>70</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>25</td>
<td>Living Environment Comment</td>
<td>27037</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>26</td>
<td>Living environment comments</td>
<td>650</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>27</td>
<td>Living environment comments</td>
<td>100</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>28</td>
<td>Living Environment Concerns</td>
<td>992</td>
<td>VLS</td>
<td>**</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>29</td>
<td>Living Status</td>
<td>35</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>30</td>
<td>RETIRED Living environment comment</td>
<td>43</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>31</td>
<td>Role/Living Environment Comments</td>
<td>319</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>32</td>
<td>Temporary Family Living Arrangements</td>
<td>6035</td>
<td>VLM</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>33</td>
<td>Type of Residence</td>
<td>68366</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>34</td>
<td>Where Is Home For Me?</td>
<td>1437</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>35</td>
<td>Where Is Home For Me?</td>
<td>336</td>
<td>FT</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>36</td>
<td>Residence-Bed Bath Location</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>Additional Bathroom Location/Comments</td>
<td>61</td>
<td>FT</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>38</td>
<td>Additional Bathroom Set Up/Equipment</td>
<td>20</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>39</td>
<td>Primary Bathroom Location/Comments</td>
<td>266</td>
<td>FT</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>40</td>
<td>Primary Bathroom Set Up/Equipment</td>
<td>219</td>
<td>VLM</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>41</td>
<td>Residence-Facility Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>Assisted Living Facility</td>
<td>446</td>
<td>FT</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>43</td>
<td>Care Facility Name</td>
<td>701</td>
<td>FT</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>44</td>
<td>Group Home</td>
<td>170</td>
<td>FT</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>45</td>
<td>Skilled Nursing Facility</td>
<td>48</td>
<td>VLS</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>46</td>
<td>Skilled Nursing Facility</td>
<td>5065</td>
<td>FT</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3. Unique flowsheet rows with observation counts and topic areas. Obs Count: number of observations for the row in the final dataset. Row Format (Data Type): FT=free-text, VLM=value list multi-select, VLS=value list single-select. Topic Areas: RS=Residence, LS=Living Situation, LC=Living Conditions. **Value lists contained only “yes” and “no” values with free-text comments.
There was also inconsistent terminology used to label the free dwelling in which the patient lived.

- 53 RETIRED # of Stairs to enter home: 4990 VLS X - - X - - 1
- 54 RETIRED # of stairs within home: 5261 FT X - - X - - 1
- 55 Stair Railings: 17 VLM X - - - - - 1
- 56 Stair Railings at Home: 30 VLM X - - X - - 1
- 57 Stair Railings at Home (VLM): 13356 VLM X - - X - - 1
- 58 Stair Railings Entering Home: 13 VLM X - - - - - 1
- 59 Stair Railings Inside Home: 3 VLM X - - - - - 1

### Residence-Safety
- 60 Current assistive devices: 581 VLM X - - X - - 1
- 61 Equipment Used at Home: 47806 VLM X - - X - - 1
- 62 Home Accessibility: 41458 VLM X X - X X - 2
- 63 Home Accessibility Comments: 26 FT X - - - - - 1
- 64 Home Safety Comments: 35 FT X X X - - - 3
- 65 Home/community accessibility: 1433 FT X - - X - - 1
- 66 Home/community Accessibility Comments: 163 FT X X - - - 2
- 67 Home/community Accessibility Comments: 30 FT - X - - - - 1
- 68 Home/community Accessibility Comments: 1 FT X - - - - - 1
- 69 Home/community Accessibility Comments: 3 FT X X - - - 2
- 70 Home/community Accessibility Comments: 3 FT - X - - - - 1
- 71 Home/community Accessibility Comments: 16 FT - X - - - 1
- 72 Home/community Accessibility Comments: 877 FT X X - X - 2
- 73 Home/community Accessibility Comments: 41 FT X X - - - 2
- 74 Home/community Accessibility Comments: 246 FT X X X - - 3
- 75 Home/community Accessibility Comments: 24 VLS - X - - - 1

### Living Conditions
- 76 Family Pets: 1427 FT - - X - - - 1
- 77 Family Pets: 325 FT - - X - - - 1
- 78 Pets? (Name/type): 425 FT - - X - - - 1
- 79 Potentially Unsafe Housing Conditions: 26 VLM - X - - X 1

### Living Situation
- 80 Current Community Support: 2478 VLM - X - X X 3
- 81 Current living arrangement: 1 VLM - X - - - 1
- 82 Lives With: 724 VLM - X - X X 2
- 83 Lives With: 10 VLM - X - X X 2
- 84 Lives With: 349 VLM - X - X X 2
- 85 Lives With: 112 VLM - X - - X 1
- 86 Lives With: 173656 VLM - X - X X 2
- 87 Lives With: 16176 VLM - X - - X 1
- 88 Lives With: 1072 VLM - X - X X 2
- 89 Lives With: 13 VLM - X - - - 1
- 90 Living Situation prior to adoption: 50 VLM X - - - - 1
- 91 Others Living In Residence (Non-Familial Relationships): 139 FT - X - - X 1

**Totals**: 604,616

Of the total 91 rows, 46% (42) were built for free-text documentation, 45% (41) were built with value lists that allowed for multiple values to be selected from the value list, and 9% (8) were built with value lists that allowed only one value to be selected from the list. Of the 49 rows with value lists, 34 had value lists that contained values related to *Residence*, 13 related to *Living Situation*, and 1 had a value list related to *Living Conditions*. Overall, 53% (48) of the 91 rows contained data related to a single topic area, 34% (31) contained data related to two topic areas, and 13% (12) contained data related to all three topic areas. Of the 42 rows with value lists, 62% (26) contained data for a single topic area, 21% (9) two topic areas, and 17% (7) to all three topic areas. Of the 49 rows that were built to collect free-text, 45% (22) contained data related to a single topic area, 45% (22) related to 2 topic areas, and 10% (5) contained data related to all three topic areas. Lastly there were two rows “Living Environment Concerns” and “Living Arrangement Concerns” that had values of “YES” and “NO”.

In flowsheet rows that were built with the intention of collecting the same information or concepts, the row display names and the value lists were not consistent across those rows. Table 4 shows an example of three flowsheet rows that had different display names and different value list members but all three were intended to capture the type of dwelling in which the patient lived.

There was also inconsistent terminology used to label the free-text rows. Comparing rows with the same display name there was variability in what data the row was intended to collect. For example “Living Environment” rows
that in most cases were intended to document residence type with value lists but in some cases the row was intended to document stairs and railings outside and inside the home. Subsequently, the measure values from the 49 rows with value lists were compiled into separate lists for each of the related elements (Table 5).

Table 4. Row display name and value lists for rows documenting the type of physical residence.

<table>
<thead>
<tr>
<th>Row Display Name</th>
<th>Type of Residence</th>
<th>Living Arrangements</th>
<th>Living Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Values in value list</td>
<td>Assisted Living</td>
<td>Apartment</td>
<td>Apartment/condo</td>
</tr>
<tr>
<td>Group Home</td>
<td>assisted living</td>
<td>Home Care Staff</td>
<td>condominium</td>
</tr>
<tr>
<td>Homeless</td>
<td>correctional facility</td>
<td>Nursing Home</td>
<td>extended care facility</td>
</tr>
<tr>
<td>Private Residence</td>
<td>foster care</td>
<td>Other (Comment)</td>
<td>group home</td>
</tr>
</tbody>
</table>

Table 5. Flowsheet measure value list compilation (raw data) categorized in the three topic areas.

<table>
<thead>
<tr>
<th>Residence Type</th>
<th>Residence Stairs &amp; Railings</th>
<th>Residence Safety (Assistive devices)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult Foster Care</td>
<td>Entering home</td>
<td>Bath bench</td>
</tr>
<tr>
<td>Apartment</td>
<td>entered home</td>
<td>Bath not on first floor</td>
</tr>
<tr>
<td>Apartment/condo</td>
<td>inside home</td>
<td>Bed and bath are not on the first floor</td>
</tr>
<tr>
<td>Assisted Living</td>
<td>inside, present at both sides</td>
<td>Bed not on first floor</td>
</tr>
<tr>
<td>Condominium</td>
<td>inside, present on left side</td>
<td>Extended tub bench</td>
</tr>
<tr>
<td>Correctional facility</td>
<td>inside, present on right side</td>
<td>Grab bars present (bathtub)</td>
</tr>
<tr>
<td>Foster care</td>
<td>none inside home</td>
<td>Hand held shower</td>
</tr>
<tr>
<td>Group Home</td>
<td>other (must comment)</td>
<td>Raised toilet</td>
</tr>
<tr>
<td>Home</td>
<td>outside, present at both sides</td>
<td>Ramps present at home</td>
</tr>
<tr>
<td>Home Care Staff</td>
<td>outside, present on left side</td>
<td>Shower grab bar</td>
</tr>
<tr>
<td>Homeless</td>
<td>outside, present on right side</td>
<td>Shower stall</td>
</tr>
<tr>
<td>Hospitality house</td>
<td>present at both sides</td>
<td>Toilet</td>
</tr>
<tr>
<td>Hotel</td>
<td>present of both sides</td>
<td>Toilet grab bar</td>
</tr>
<tr>
<td>Hotel/motel</td>
<td>present on left side</td>
<td>Tub/shower combo</td>
</tr>
<tr>
<td></td>
<td>stairs (1 railing present)</td>
<td>Tub/shower is not walk in</td>
</tr>
<tr>
<td></td>
<td>stairs within home</td>
<td></td>
</tr>
<tr>
<td></td>
<td>entering home</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Living Situation Subject</th>
<th>Living Conditions Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adoptive parent(s)</td>
<td>Chipped paint</td>
</tr>
<tr>
<td>Alone</td>
<td>No running water</td>
</tr>
<tr>
<td>Aunt</td>
<td>Insects/pests</td>
</tr>
<tr>
<td>Birth family</td>
<td>No air conditioning</td>
</tr>
<tr>
<td>Brother</td>
<td>No electricity</td>
</tr>
<tr>
<td>Child(ren)</td>
<td>No heat</td>
</tr>
<tr>
<td>Child(ren), adult</td>
<td>Paint chipped</td>
</tr>
<tr>
<td>Child(ren), dependent</td>
<td>No hot water</td>
</tr>
<tr>
<td>Domestic partner</td>
<td>Stairwells unsafe</td>
</tr>
<tr>
<td>Father</td>
<td>No indoor plumbing</td>
</tr>
<tr>
<td>Father and partner</td>
<td>Unable to assess</td>
</tr>
<tr>
<td>Foster care</td>
<td>No lighting</td>
</tr>
<tr>
<td>Foster family</td>
<td>Unsafe stairways</td>
</tr>
<tr>
<td>Foster parent(s)</td>
<td>No phone</td>
</tr>
<tr>
<td>Friend(s)</td>
<td>Windows broken</td>
</tr>
<tr>
<td>Grandchild(ren)</td>
<td></td>
</tr>
</tbody>
</table>
Groupings generated five concepts: (1) Residence Type, (2) Residence Stairs and Railings, (3) Resident Safety, (4) Living Situation Subject, and (5) Living Conditions Type. The value lists showed variability in terms used as well as terms that did not necessarily belong (e.g., names of facilities like “Ronald McDonald House”).

Lastly, differences were found in the format of a number of the rows collecting similar data. For example, five unique rows were found that were intended to collect the names and contact information for care facilities each one specific to the facility type. Four of the five flowsheet rows were built as free-text measure values. Two of the rows had the same display name but one was a free-text measure value and the other was built as a custom list.

**Discussion**

SDOH play an important role in the provision of care. However, EHRs, for the most part, often do not contain well-designed documentation tools to collect these data or standardized storage of this information for clinical care and secondary uses. In our analysis of a large healthcare enterprise and three key SDOH, we observed that much of this information is documented in free-text form in notes or in flowsheets rows, the latter of which was the focus of this study. Flowsheets allow for documentation in a longitudinal manner; however, their format is really most efficient for short answer or numeric values. Their ongoing management across settings and certain types of care remains an important challenge. This study demonstrates that flowsheet tools are being used to document SDOH. However, the overall design of these tools is not optimal and the use of any specific standard or terminology is not clearly evident, which has broad implications to interoperability and secondary use of these data.

In total, 91 rows were found to contain either measures and/or free-text comments related to the three topic areas. Of the 49 rows built with value lists, there was significant duplication in rows with regards to naming and also multiple rows seemingly built to collect the same concepts. For example, 22 unique rows were related to Residence and were built with value lists to collect the type of dwelling in which the patient lives. We found inconsistencies in terminology used to label rows as well as inconsistencies in terminology and content in value lists. For example, Residence Type was documented in rows that were labeled “Type of Residence”, “Living Arrangements”, and “Living Environment”. This duplication and inconsistency demonstrates a need for more rigorous content and knowledge management of EHR documentation tools as well as a need for more attention towards overall design and architecture.

The free-text comments in many cases did not match the intent of the flowsheet row as determined by review of the value lists associated with the row. For example, the row “Type of Residence”, which was built as a multi-select value list, had free-text comments that included the name of the facility, who was living with the patient in that residence, location such as addresses, and information related to residence details such as stories and stairs. In another row entitled “Lives with”, the free-text comments included information such as the facility name, dwelling type, pets (numbers and names), details about where other family members live, and number of children and their ages.

A comparison of the flowsheet value lists compiled from this work (Table 5) with the value lists obtained from previous standards work demonstrated overall similar values for residence type and living conditions although some standards, most prominently SNOMED CT, were more comprehensive in these topic areas. However, we found differences in details specifically values related to stairs, railings, and installed safety devices. The amount of flowsheet documentation found regarding these details, could indicate the importance and relevance of this information for patient care and therefore should be included in future comprehensive standards development work.

Another issue found was excessively long value lists that were high maintenance. The row “Skilled Nursing Facility” had a value list that contained well over 100 items. The value list items were composed of residential facility names, locations, and phone and fax numbers. Flowsheet metadata for these rows indicated that these value lists are updated frequently as facility names or phone and fax numbers change thus this model requires more maintenance to stay current. This brings to question whether a flowsheet is the optimal location in the EHR for a patients place of residence. The EHR does contain a designated place for patients address in the demographics section; however, many times for patients who live in residential facilities, the address entered into the patient demographics may be that of a relative or guardian. Also, since flowsheets are not necessarily readily available to all providers, it may be better to have residential facility information located elsewhere in the EHR such as in the demographic sections of the chart where it is more readily accessible for viewing and updates but separate from the billing address.

Next steps for this work include compiling and harmonizing the flowsheet value lists with the value lists from previous work evaluating existing standards and EHR free-text documentation. In addition, formal annotation
techniques could be used to enhance natural language processing tools for these topic areas. Since the ultimate goal is to update the model representation from previous work, that modeling should be done using formal model representation tools such as openEHR, as well as aligning the model with existing standards such as Clinical Information Modeling Initiative and terminologies such as the Omaha System\textsuperscript{56} for the three topic areas of \textit{Residence, Living Situation, and Living Conditions}. Lastly, future work could also include an examination of patient level SDOH versus environmental level SDOH and impact to health outcomes.

**Conclusion**

In summary, this study demonstrated the wide variation in the design and use of flowsheet rows as data collection tools for SDOH information related to \textit{Residence, Living Situation, and Living Conditions}. In addition, data quality is less than optimal due to the lack of standards in terminology for the element names and the value lists as well as the extensive use of free-text measure values and comments. As a result, there remains an opportunity to redesign these flowsheet rows to create efficiencies in documentation and to optimize the quality of the data being collected through pruning and combining of similar concepts into fewer rows and developing comprehensive standards and utilizing coding systems for these data. Overall, the lack of standards for SDOH documentation has implications for interoperability and secondary use of data.

**Acknowledgements**

This work was supported by the National Library of Medicine of the National Institutes of Health (R01LM011364) and University of Minnesota Clinical and Translational Science Award (8UL1TR000114-02). The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

**References**


CHCi – A Dynamic Data Platform for Clinical Data Capture and Use

Danny T.Y. Wu, PhD, MS1,2,3 Kai Zheng, PhD4,1,5, David J. Bradley, MD2
1School of Information, University of Michigan, Ann Arbor, MI; 2Division of Pediatric Cardiology, University of Michigan, Ann Arbor, MI; 3Department of Biomedical Informatics, University of Cincinnati, Cincinnati, OH; 4Department of Health Management and Policy, University of Michigan, Ann Arbor, MI; 5Department of Informatics, University of California, Irvine, CA

Abstract

All academic medical centers have a strong desire to maximize the value of their clinical data for secondary use purposes such as quality improvement (QI) and research. However, this need has not been adequately fulfilled due in part to the fact that the data capture functions in current electronic health record systems predominantly focus on clinical documentation and billing, lacking the flexibility to allow the collection of additional data elements critical to QI or research. To address this gap, we designed and developed a dynamic data platform to support clinicians’ varied needs for recording additional data about their patients outside of direct patient care (e.g. classifying patient conditions based on the inclusion criteria of a research-oriented patient registry). In this paper, we describe the design considerations of this platform such as data models, query functions, coding and controlled vocabulary, user interface design, access control, and data interoperability. In developing the platform, we partnered with the frontline clinicians in an academic congenital heart center, and adopted the agile software development approach with numerous rounds of evaluation and iterative refinement. Since 2013, this platform has been successfully used to meet the dynamic QI and research data needs of clinicians in the congenital heart center. Future work includes improving the efficiency and effectiveness of the platform and incorporating cutting-edge data interoperability standards.

Introduction

Academic medical centers are dedicated to providing high-quality patient care and conducting clinical and translational research (1,2). To achieve these goals, it is essential to ensure accurate and comprehensive recording of clinical data. For example, without such data, the promise of genome-wide association studies and comparative effectiveness research cannot be adequately fulfilled. This critical need, however, has not been fully met by current electronic health record (EHR) systems, as the priority in the design of these systems is to support direct patient care as well as the administrative processes (e.g., billing). They do not provide much flexibility to allow for the collection of extra data elements that are crucial for secondary use purposes such as quality improvement (QI) and research.

To address this gap, there is a need to extend current EHR systems’ capability so that clinicians can, as part of their patient care workflow, record additional information about their patients or clinical processes that can be later used to facilitate QI and research (3,4). Currently, there have been initiatives that utilize application programming interfaces (APIs) or other means of integration to build research-oriented data collection into EHRs (e.g. integration with clinical trial management systems to support point-of-care patient recruitment). However, these solutions may be costly, cumbersome (e.g. 5), require a significant amount of vendor cooperations, and are difficult to maintain due to frequent vendor software upgrades and/or legal regulatory compliance. Therefore, these solutions do not fully address the changing data needs for supporting QI or research.

A potential solution is to develop companion software systems that work side-by-side with an EHR to provide supplementary functionality in supporting research data capture. There has been considerable effort in developing and adopting such IT solutions in healthcare institutions. However, existing solutions are either designed as standalone tools and are therefore difficult to integrate into clinicians’ workflow, or are largely redundant with EHRs, requiring a significant amount of double data entry.

In this paper, we describe a dynamic clinical data platform that we designed and developed to address this gap. Prior to deploying this data platform, many database applications, using technologies such as Access databases, existed at a leading academic congenital heart center to support clinicians’ data needs. However, these legacy database applications were implemented in an ad-hoc manner, resulting in common health IT problems such as duplicate programming effort and silo datasets. This was the main motivation for us to develop a unified platform that are highly modularized and customizable, and able to integrate discrete datasets together to facilitate downstream data uses to maximize the value of clinical data. This dynamic data platform has four design objectives: 1) high extensibility to
address clinicians’ dynamic data needs in a timely fashion, 2) high accessibility to the data stored on the platform, whether structured or unstructured, 3) high accountability to justify the use of data records through detailed audit trail logs, and 4) high interoperability with other health IT applications, especially the EHR, to facilitate data sharing and minimize redundant data entry. In this manuscript, we report the design considerations of the platform, and discusses the value of this dynamic data platform through empirical evidence derived from the implementation of the system in a leading academic congenital heart center.

Methods

Clinical Setting

The design and implementation of the dynamic data platform were made possible through the collaboration of the Michigan Congenital Heart Center (MCHC), a multi-specialty inpatient and outpatient cardiac center where pediatric cardiologists, pediatric cardiac surgeons, certified nurses, and technicians work together to treat patients with pediatric and congenital heart diseases. Due to its size and multi-specialty nature, the need for clinical data capture and use in the center has become a more imperative issue than other clinical settings. For example, the MCHC catheterization team used a commercial case management tool to support its procedural workflow. This tool, however, shared no information with the institutional EHR system or with other MCHC teams, although it contained a potentially valuable dataset that could have been better utilized for quality improvement (e.g. complications review) and clinical research.

System Design and Evaluation

The implementation of the data platform in MCHC is named the “Congenital Heart Center intelligent and innovative data platform” (referred to as “CHCi” hereafter), which began in mid 2013. The database team has been working in the clinical setting since 2011, accumulating two years of experience prior to the implementation of CHCi. In these two years, the database team adopted traditional software development method (waterfall model) and standard relational database and web development techniques, producing four separate functional database applications to support four MCHC subgroups. A formal evaluation was conducted by a group of usability experts in the beginning of 2013 to understand the effectiveness, usability, and user acceptance of the four database applications, using human-computer interaction system evaluation methods including interactive map, comparative analysis, interview, survey, heuristic evaluation, and usability testing. The evaluation results suggest improvements on entry error prevention, navigation support, user interface design, and search and sorting functions. This formal evaluation led to the redesign of the software architecture and the adoption of agile software development method to create CHCi in mid 2013. Since then, the database team has been iteratively refining CHCi to enhance its ability to support the missions of many subgroups in MCHC using both top-down and bottom-up approaches.

After three years of use, CHCi has achieved many successes as described in our previous work (6). The current paper aims to share the design considerations and implementation challenges of this platform. One key design challenge was that each MCHC subgroup has both shared and unique data needs. CHCi was therefore designed to host multiple database applications for each of the subgroups. These related database applications were termed “modules” of the platform. The platform and its modules were designed such that a function of one module can be shared with, or easily reused by, another module. Similarly, the datasets captured in one module can be easily connected to the databases underlying the other modules. In the following section, we describe the approaches to address this design challenge and share technical details of CHCi with regard to the data model, controlled diagnostic codes, data search, user interface design, platform interoperability, and data access control. In terms of evaluation, we have demonstrated several successful case studies and preliminary evaluation results in our previous publications (7–9). In this paper, we provide another evidence to show the high extensibility of CHCi, which is measured by how many new modules were created and used between 2013 and 2016 versus the four data base applications implemented in the previous two years. We also provide informal evaluation evidence to show the high accessibility, accountability, and interoperability of CHCi. These evaluation results are described at the end of the results section.

Results

Data Model

An Entity-Attribute-Value (EAV) data model was chosen to support the storage and access of heterogeneous data on the CHCi platform (10). Table 1 shows an example of a traditional table and its corresponding EAV model. In this example, event attributes including event status and doctor of record in the left table as separate columns are transformed into key-value pairs in the right table. This decision was made because most current health IT systems are still supported by relational databases. Deploying an EAV model within a relational database, rather than using
schema-less databases, could lower the barriers to integrating platform data with existing data tables, resulting in higher interoperability of the platform. Utilizing such EAV data model can also favorably impact our two other design objectives: 1) high extensibility: adding an attribute does not change the table schema and the corresponding query functions; 2) high accessibility: data records can be easily queried against a smaller set of columns with limited or no effort to join multiple tables to obtain information (11).

Table 1. An example of a traditional table (left) and its corresponding Entity-Attribute-Value (EAV) data model (right)

<table>
<thead>
<tr>
<th>MRN*</th>
<th>CASE_ID</th>
<th>STATUS</th>
<th>DOCTOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>01</td>
<td>Admitted</td>
<td>Bradley</td>
</tr>
<tr>
<td>0001</td>
<td>02</td>
<td>Admitted</td>
<td>Wu</td>
</tr>
<tr>
<td>0002</td>
<td>03</td>
<td>Discharged</td>
<td>Bradley</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SEQ</th>
<th>MRN</th>
<th>CASE_ID</th>
<th>ATTRIBUTE</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0001</td>
<td>01</td>
<td>STATUS</td>
<td>Admitted</td>
</tr>
<tr>
<td>2</td>
<td>0001</td>
<td>01</td>
<td>DOCTOR</td>
<td>Bradley</td>
</tr>
<tr>
<td>3</td>
<td>0001</td>
<td>02</td>
<td>STATUS</td>
<td>Admitted</td>
</tr>
<tr>
<td>4</td>
<td>0001</td>
<td>02</td>
<td>DOCTOR</td>
<td>Wu</td>
</tr>
<tr>
<td>5</td>
<td>0002</td>
<td>03</td>
<td>STATUS</td>
<td>Discharged</td>
</tr>
<tr>
<td>6</td>
<td>0002</td>
<td>03</td>
<td>DOCTOR</td>
<td>Bradley</td>
</tr>
</tbody>
</table>

* Medical Record Number

Moreover, the EAV model in our platform is expanded to achieve high accountability and modularization. With additional user identifiers and timestamps, our EAV model keeps all the “transactions” of data entries and can effortlessly generate audit trail logs and data entry trajectories, leading to higher data accountability since no data records are deleted on the platform.

Figure 1 – CHCi EAV Data Model (the links indicate the view generation process, not table relationships).

In terms of modularization, each module on the platform is supported by exactly the same six tables, including the EAV model and five supporting tables and views. As shown in Figure 1, a module contains: 1) a data dictionary storing user-defined variables, 2) a code table storing discrete options for the variables, 3) a transaction table (EAV model) that keeps all the data entries, 4) a view combining the transaction table with the code table to facilitate data searches, and 5,6) two additional views to “pivot” data in the EAV model to a traditional table schema to support data access and downstream uses. These pivot tables are necessary because EAV models are not very human-readable and cannot be easily manipulated or combined with other data for analytics. Having the same table schema across all modules unifies the module creation process on our data platform. Moreover, this model shortens the development cycle by reusing the same data manipulation functions (e.g. UPDATE and ARCHIVE) across all modules. It is worth noting that all the values stored in our EAV model are converted into strings no matter what their original data type is. The original data types are specified separately in the data dictionary. This design allows all data entry records to be stored in one EAV model to simplify data storing and manipulating processes.
Controlled Diagnostic Codes

In the data model, all data variables can be flexibly defined by module users except diagnostic codes due to the unique value provided by diagnostic coding in clinical work and research. A special diagnostic code table is created and maintained by the database team, in particular under a centralized control of the lead physician. Module users, however, can create a temporary code if they feel the current code book is lacking, along with a code request to the lead physician. These code requests are carefully reviewed by the lead physician and the decision is made either to creating a new code, modify an existing one, or assign an existing code to the requested term. Temporary codes are automatically converted into permanent ones based on these decisions. This design has the benefit of ensuring the uniformity and usefulness of the diagnostic codes on our platform.

Another design consideration related to diagnostic codes is the adoption of Interface Terminology, the idea of using locally agreed terms by clinicians to facilitate clinical data entry, rather than using standardized terminology such as SNOMED CT or ICD9 codes (12). To promote downstream data uses, interface terminology can be later converted to standardized terminology. Figure 2 shows an example of our diagnostic code component shared across all CHCi modules. In this example, a term “down” is searched for “Down’s Syndrome” and a list of possible diagnostic codes is suggested by the data platform. The searcher picks the code “14.01.02” with a formal description “Trisomy 21”, which is automatically mapped to standardized terminology such as the International Paediatric Congenital Cardiac Code (IPCCC) and the Society for Thoracic Surgeons (STS) nomenclature frequently used in Pediatric Cardiology.

Figure 2 – Searching “Down’s Syndrome” [top right] in the Diagnostic Code Component (the results rank “Trisomy 21” the top choice [bottom left], which is mapped to other standard codes [bottom right])

Data Search

Our data model accommodates both structured and unstructured clinical data. Here, unstructured data refer to free-text clinical documents such as progress notes and discharge summaries. These notes are exported from our institutional EHR system’s data repository as plain text files and their note identifier is stored in the transaction table (EAV data module) as a key-value pair. The text files are further indexed and maintained by Apache Lucene (a full-featured text search engine library) to allow basic information retrieval. Figure 3 shows the general search process of the platform. Take the search term “smoker” for example, this keyword is first queried against the EAV data model and its views to identify records with “smoker” as part of the title and the value string. Next, this keyword is searched against the text indexes to retrieve notes with matched keywords. Then, these two results are combined at the case level, and eligible patient cases are presented with matched variables shown in context.

By combining the records in our EAV model and the indexes in Apache Lucene, this design can fully utilize the benefits of both structure and unstructured data and therefore decreases the tension between these two data types. As Rosenbloom et al. indicate, this tension comes from the complementary nature of the two data types. While structured data have the benefits of standardization and computer analyzable format, unstructured data have higher expressivity, flexibility, and may be more suitable for workflow (13). The ability of our data platform to search both structured and unstructured data can lower this tension and increase the utilization of clinical data regardless of their data types.
User Interface Design

The user interface is separated in two levels to help module users navigate through the data to accomplish their tasks. At the first level, the interface is designed to provide an “overview” of case events along with various ways to access these events and to export them. This interface consists of a landing page listing patient cases in a reverse chronological order by default (Figure 4). Module users can then browse case events using the paging options, search cases by entering keywords in the general search bar, filter cases based on predefined values in the advance search panel, and save search results in the CSV format by clicking the export button. At the second level, the user interface is designed to support data entry and verification. This second-level edit page shows patient demographics on top with additional information (e.g., gender, date of birth, and age) in the same order as the inpatient wrist band, followed by all cases of this patient to form a medical history (trajectory) so that a case can be interpreted in the context of adjacent ones (Figure 5). All the data entry activities on this edit page are automatically recorded using asynchronous JavaScripts (jQuery) to prevent data loss in a busy clinical environment. To facilitate case browsing, the content of a case can be collapsed and expanded by clicking its header. In addition, audit trail logs of data entries can be easily reviewed by clicking the “Hx” button to the right of the case header. This can strengthen module users’ trust in data integrity of the platform.
This two-level user interface design nicely balances the need for uniformity and customization of the CHCi modules. Since the first-level landing page can be shared across modules, it maintains the same look-and-feel and minimizes implementation effort when creating a new module. This landing page can also be configured to show different columns and identifying icons for each module through a database table. On the other hand, the second-level edit page is highly customizable. It can present data elements in various layout components (e.g. dropdown, radio button, text box, etc) and equip them with interactive features (hide/show, auto-calculation, data validation, etc). This high customizability results from the modularization of user interface components and the linking between these interface components and the data dictionaries. Taking dropdown creation for example, a variable is first entered in the data dictionary with its options being listed in the corresponding code table. Then, a dropdown interface component (programming function) is called, with parameters indicating the variable in the data dictionary to create a usable dropdown on the edit page. This dropdown can be easily repositioned on the edit page based on users’ preferences. If
the dropdown needs complicated interactive features, these features can be implemented in JavaScript as a parameter of the interface component.

**Platform Interoperability**

The data platform currently has three major data interfaces, namely 1) web services, 2) direct database connections, and 3) file transfer. The choice of data interfaces depends on the target data sources and the timeliness of data needs. Figure 6 illustrates the conceptual model of the platform interoperability. For connections to the institution’s EHR system, the platform employs XML-based web services if real-time data access is necessary. Otherwise, the platform retrieves data from the EHR data repository through a direct database link. For connections to other health IT applications, the data platform either uses a direct database link if permitted, or receives data files from those applications. These connecting processes are automated and run in batches in various intervals, between daily to monthly, depending on the module needs. Once the data platform collects all desired clinical data from heterogeneous sources, module users can interact with them through the user interfaces as described in the previous section.

![Diagram of CHCi Interoperability](image)

**Figure 6 – Conceptual Model of CHCi Interoperability**

**Data access control**

The data platform consists of a web server and a database server maintained by the University of Michigan Medical School Information Technology team. These two servers are hosted behind the Medical School firewall and cannot be accessed outside of the network without using Virtual Private Network (VPN). Even with a VPN and Medical School credentials, users must be granted access to the platform to link their credential to the platform login. Moreover, platform users can be granted different levels of access depending on their needs and job responsibilities. The types of access include 1) read-only, 2) read and write, and 3) full (read-write-verify) access. Data access is module specific, meaning that a user can be assigned read-only access to one module and full access to another. In addition to accessing data through the user interface, users can submit data requests to the database team. However, each research-oriented data request must have its own approval from the Institutional Review Board. Data requests are carefully reviewed and approved by the lead physician before being executed by the database team.

**Implementation**

The dynamic data platform (CHCi) was developed between January and June 2013, and has been running in the congenital heart center since September 2013. The platform is implemented in ASP.NET MVC 4.5 and hosted on a Microsoft IIS 7.5 web server. MVC (Model-View-Controller) is a software design and development framework that separates data from their business logic and representation to create a cleaner software architecture. The extended
EAV data model is implemented in an Oracle Database 12c Enterprise Edition 64-bit. Clinical notes are stored as plain text files and indexed using Lucene.Net 3.0.3. The user interfaces are programmed using HTML5 and JavaScript libraries including JQuery 2.0 and Twitter Bootstrap 2.3.2.

Evaluation

The high extensibility of the dynamic data platform is supported by the fact that the database team has been able to fast prototype new modules and iteratively refine them. Prior to using this data platform, between July 2011 and August 2013 (26 months), the database team deployed only four modules in an ad-hoc manner for the magnetic resonance imaging (MRI), neurodevelopmental follow-up (ND), ambulatory cardiac recording (HOLTER), and fetal cardiology (FETAL) teams to support their clinical work and research needs. After using CHCi, between September 2013 and December 2016 (40 months), the database team not only migrated the four legacy modules to CHCi, but also added nine more fully functional modules. Table 1 lists all CHCi modules as of 2016, their status, and the number of users. The total number of users is not the sum of all module users because one user may play multiple roles in different modules and can have a different level of access.

Table 1 – Current modules on the dynamic data platform

<table>
<thead>
<tr>
<th>Module ID</th>
<th>Description</th>
<th>Status*</th>
<th># Users**</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRI</td>
<td>Magnetic resonance imaging</td>
<td>M</td>
<td>2</td>
</tr>
<tr>
<td>ND</td>
<td>Neurodevelopmental follow-up</td>
<td>M</td>
<td>3</td>
</tr>
<tr>
<td>HOLTER</td>
<td>Holter (ambulatory electrocardiography)</td>
<td>M</td>
<td>8</td>
</tr>
<tr>
<td>FETAL</td>
<td>Fetal cardiology</td>
<td>M</td>
<td>4</td>
</tr>
<tr>
<td>CATH</td>
<td>Cardiac catheterization</td>
<td>M</td>
<td>5</td>
</tr>
<tr>
<td>LONG</td>
<td>Longitudinal follow-up</td>
<td>M</td>
<td>4</td>
</tr>
<tr>
<td>PC4</td>
<td>Pediatric Cardiac Critical Care Consortium – Intensive care</td>
<td>M</td>
<td>4</td>
</tr>
<tr>
<td>ECMO</td>
<td>Extracorporeal membrane oxygenation</td>
<td>U</td>
<td>7</td>
</tr>
<tr>
<td>ECHO</td>
<td>Echocardiography</td>
<td>U</td>
<td>4</td>
</tr>
<tr>
<td>NOTE</td>
<td>Clinical notes</td>
<td>U</td>
<td>10</td>
</tr>
<tr>
<td>ACHD</td>
<td>Adult congenital heart Disease</td>
<td>U</td>
<td>5</td>
</tr>
<tr>
<td>EXER</td>
<td>Exercise Laboratory</td>
<td>D</td>
<td>8</td>
</tr>
<tr>
<td>OHT</td>
<td>Transplant service</td>
<td>D</td>
<td>3</td>
</tr>
</tbody>
</table>

* M (Maintenance), U (User Testing), D (Development)
** Physicians, nurses, technicians, project coordinators, and data analysts

The high accessibility of the platform is attested by the frequent user-generated search and data export on the platform. For example, users exported about 202 and 174 datasets in calendar years 2015 and 2016, respectively, in addition to ad-hoc data requests that directly pull records from the database. While the high accountability of the platform is not documented systematically, the database team did occasionally encounter situations where module users needed to revert the current value of a variable to a previously entered value. Since our EAV model does not delete any records, this type of request is very easy to handle, which increases the accountability of the platform. Lastly, the interoperability of the platform, especially with the institution’s EHR system, is highly appreciated by users. For example, based on our pilot evaluation, bringing all information to a centralized place for the Holter users as a result of high interoperability likely contributed to the substantial improvement in case turnaround time (7).
Although this implementation has shown the achievement of our four design objectives, it also faced a potential scalability challenge. That is, when a module had too many records, it did not respond to a search query very quickly. For example, 142,000 historical ECHO cases with 12 variables of each were imported to our platform, leading to 1.75 million records in the EAV data model. Querying these pivoted tables to complete data search became slow. The root cause of this issue had been identified, including: 1) the suboptimal design of the SQL statement of the pivoted views, 2) insufficient indexes on the clinical notes, and 3) inefficiency of the query process. Two solutions have been implemented to resolve this scalability issue. First, to enhance the EVA data model, the CHCi modules with a large amount of records have extra physical data tables with exactly the same table schema as the pivoted views to avoid suboptimal queries. A set of functions have been implemented to ensure consistency between the data stored in the EAV data model and these new physical data tables. Second, the indexes of clinical notes were rebuilt to include more metadata of the notes to support record filtering in the query process. These two solutions have successfully shortened the search time in the ECHO module, and have been applied to other modules facing similar challenges to improve the overall scalability of the platform.

Discussion

We designed and developed a dynamic clinical data platform (CHCi) with high extensibility, accessibility, accountability, and interoperability. Since 2013, CHCi has been successfully supporting clinicians’ variety of data needs in the congenital heart center. Our data platform can collect and harmonize clinical data that were otherwise scattered and difficult to utilize. Here, we share three lessons learned in this journey. First, modularization is the key to our success. Since the data platform is highly modularized in both front-end (i.e. user interface) and back-end (i.e. data models), the database team can rapidly respond to clinicians’ data needs (i.e. fast prototyping) so that more resources can be devoted to improving data collection workflow, increasing user satisfaction, and conducting evaluation studies to ensure the effectiveness and efficiency of the platform (7–9). Also, the high modularization nature allows the database team members to cover each other’s work when necessary and therefore makes it much easier to maintain the data platform. The second lesson we learned is the necessity of iteratively improving our modules since users’ data needs are varied and may evolve overtime. The database team works closely with the end users to ensure that the platform meets their data needs and to identify future improvement opportunities. The third lesson is that all stakeholders have come to appreciate that a significant amount of effort must be applied to educating the user base about new developments. This strengthens the trust among existing users and make the adoption easier for prospective users. It also garners essential organizational support to sustain this important work.

Through the implementation of this dynamic data platform, we have identified a potential salability issue in the design and had developed two solutions to address this issue as described above. We will continue improving the dynamic data platform to ensure the achievement of our four design goals and the overall efficiency and effectiveness of the platform. Meanwhile, we are enhancing the platform’s ability to support clinical research. For example, we have started to work with the research group in MCHC to develop a patient cohort management tool to facilitate the identification of study populations based on the case attributes and notes. In addition, we would like to increase the portability of the data platform to benefit other health institutions facing similar data capture issues. We have started to re-engineer the next version of the platform, automate the deployment processes, and test cutting-edge data interoperability standards, such as SMART-on-FHIR (14). We are very excited to continue this journey to develop this dynamic data platform to support clinical data capture and facilitate data use, and demonstrate the value of our platform in reducing clinicians’ data entry burdens, improving care quality, and supporting long-term registry management.

Conclusions

In this paper, we present the design considerations of our dynamic data platform, which has been implemented at an academic congenital heart center and successfully supports the clinicians’ dynamic data needs since 2013. We learned that modularization, iterative improvement, and user education are the keys to our success. We believe the platform has great potential to serve as a long-term registry management system, and continue improving the platform, enhancing its ability to discover patient cohorts, and incorporating cutting-edge data interoperability standards.

Acknowledgments

We thank Dr. Caren Goldberg, Dr. Sara Pasquali, Dr. Gerold Serwer, Mr. Ray Lowery, Ms. Janet Donohue at the Michigan Congenital Heart Center for their advice on improving the usability and usefulness of the platform. We also thank Ms. Yue (Anna) Yu for her help summarizing the achievements of the platform. This project was not supported by any grant at the time of submission. The first author worked part-time as a senior database analyst/programmer at
the Michigan Congenital Heart Center between 2011 and 2016 when he was pursuing his PhD degree at the University of Michigan School of Information, and served as a technical consultant to the project between March and October 2017 after graduation.

References

A Deep Learning Approach to Examine Ischemic ST Changes in Ambulatory ECG Recordings

Ran Xiao, PhD1, Yuan Xu, PhD1, Michele M. Pelter, RN, PhD1, David W. Mortara, PhD1, Xiao Hu, PhD1,2,3,4
1Department of Physiological Nursing, University of California, San Francisco, CA; 2Department of Neurological Surgery, University of California, San Francisco, CA; 3Institute for Computational Health Sciences, University of California, San Francisco, CA; 4Core Faculty, UCB/UCSF Graduate Group in Bioengineering, University of California, San Francisco, CA

Abstract
Patients with suspected acute coronary syndrome (ACS) are at risk of transient myocardial ischemia (TMI), which could lead to serious morbidity or even mortality. Early detection of myocardial ischemia can reduce damage to heart tissues and improve patient condition. Significant ST change in the electrocardiogram (ECG) is an important marker for detecting myocardial ischemia during the rule-out phase of potential ACS. However, current ECG monitoring software is vastly underused due to excessive false alarms. The present study aims to tackle this problem by combining a novel image-based approach with deep learning techniques to improve the detection accuracy of significant ST depression change. The obtained convolutional neural network (CNN) model yields an average area under the curve (AUC) at 89.6% from an independent testing set. At selected optimal cutoff thresholds, the proposed model yields a mean sensitivity at 84.4% while maintaining specificity at 84.9%.

1. Introduction
Patients with acute coronary syndrome (ACS) are at risk of transient myocardial ischemia (TMI), which can lead to serious medical complications. It has been found that more occurrence of myocardial infarction after admission, acute pulmonary edema and unplanned transfer from telemetry unit to the intensive care unit associated with patients with TMI compared to those without TMI1. As a critical step in identifying ACS, the early detection of myocardial ischemia helps reduce irreversible damage to heart tissues and prevent patient deterioration. Several deployed methods in detecting myocardial ischemia, such as coronary angiography and echocardiogram, are either invasive and/or resource demanding, or only able to access a brief time period, making them unsuitable for initial rule-out phase for ACS. On the other hand, continuous electrocardiography (ECG) provides an economical alternative and additional diagnostic values for early transient ischemia detection. It measures the electrophysiological activities of the heart in real time and noninvasively, together with other benefits including easy setup and long-term monitoring.

ECG is an important risk stratification tool in the immediate phase of ACS. ST (i.e. the isoelectric section in ECG waveform between J point and the beginning of T wave) elevation on the ECG is presented in up to 25% of ACS patients (i.e., ST elevation myocardial infarction (STEMI)), whereas the rest (non-ST elevation-ACS (NSTE-ACS) or unstable angina (UA)) show non-specific ECG changes2. This 75% of ACS patients is at risk for TMI, which can be detected with continuous ECG monitoring. However, current ECG monitoring software is underutilized due to excessive false alarms3. This further contributes to alarm fatigue, which is ranked as the top technology hazard in 2014 by the Emergency Care Research Institute (ECRI)4.

In contrary to current monitoring software, expert clinicians are capable of detecting true ST changes even if the ECG is moderately contaminated (i.e., motion artifact, patient movement, etc.) and are able to differentiate between ischemic and non-ischemic changes, by examining ECG waveforms screen by screen. Therefore, representing ECG tracings as images could provide valuable discriminative features about ST change. Meanwhile, the rapid developing approach of deep learning techniques, especially the convolutional neural network (CNN), has been constantly pushing the performance boundary of image recognition by computer algorithms5. A well-designed CNN model has even surpassed human benchmarks in a visual recognition challenge6. Some pioneer studies have adopted deep learning techniques in mining ECG features to tackle challenging medical problems related to the heart. In one study, CNN was adopted to detect various types of arrhythmia in ECG7. In another study, CNN was utilized to learn ECG features for screening paroxysmal atrial fibrillation patients8.
Inspired by the aforementioned studies, the present study proposes an image-based approach that transforms windowed ECG dynamics into images to train deep learning model for ST depression events. The proposed CNN model is trained from all-day ambulatory ECG recording sessions, and tested with an independent testing set to evaluate the model performance. As a preliminary study, we target to detect significant ST change as an initial step, with further discrimination between ischemic and non-ischemic ST changes as our future effort. Nevertheless, the reliable detection of ST provides accurate description of duration of ST change and its multi-lead patterns, which will be important for downstream differentiation between ischemic and non-ischemic ST episodes.

2. Methods

2.1 Data information

Data used in the present study are selected from the Long-Term ST Database (LTST database) from PhysioNet\textsuperscript{19}. The LTST database provides 20–24-hour ambulatory 2- or 3- lead ECG recording sessions, recorded with sampling frequency at 250 Hz\textsuperscript{11}. For each session, single-lead annotations are performed for ischemic and non-ischemic ST events, as well as noisy and unreadable segments. The annotation information is achieved semi-automatically with combination of computer software (SEMIA) and human experts\textsuperscript{11}. Detailed information about annotation procedures in LTST database can be found in Jager et al.’s publication\textsuperscript{11}. It is worth noting that annotation of significant ST events was based on three types of protocols by varying minimal amplitude of ST change (V<sub>min</sub>) and minimal duration of the change (T<sub>min</sub>). Subsequent analysis in the present study relies on annotation information from the protocol B (V<sub>min</sub> = 100 µV and T<sub>min</sub> = 30 s) as adopted in the original study. In this preliminary study, the first 20 sessions out of total 86 sessions in the database, which are from 20 unique patients, were selected to train our CNN model. This will ensure independent patients in the training and testing set. The following 15 sessions (see Table 1 for a full list of testing sessions), which are also from unique patients and do not contain excessive number of sudden-step ST events (<100), were selected as testing data to evaluate the performance of the proposed CNN model. The WFDB toolbox was implemented to extract ECG waveforms and annotation information from the LTST database\textsuperscript{11,12}.

2.2 Sample preparation

The continuous ECG waveform from each lead and session is firstly separated into episodes according to annotations of different events, which will later form case (significant ST changes) and control (no significant ST changes) conditions. Specifically, episodes related to significant ST change, such as ischemic ST changes and heart-rate related ST change, are grouped as the case condition, similar to previous studies\textsuperscript{13,14}. Then the control condition consists of the remaining ECG data after further exclusion of segments related to unreadable ECG, noise, and sudden-step ST changes caused by axis shift and conduction change. For case events, time points of episode start, maxima of ST change and episode end are provided, so event episodes can be readily extracted by a reference to starting and ending time points. However, only one event time point is available for noise and sudden-step ST events, while another equally important information, the event duration, is unknown. Due to such limitation in the database, 10-second segments before and after these event time points are marked and removed from control condition. The present study aims to detect ST changes via an image-based approach, so single-lead image samples are designed to contain 10-second temporal dynamics of ECG overlaid with standard ECG grid (0.04 seconds as horizontal interval and 0.1 mV as vertical interval). Examples for both case and control conditions can be found in Figure 1.

Different sample selecting schemes are implemented for training and testing sessions, respectively. ECG episodes with ST changes are much shorter in length comparing to normal ST cases (see Table 1 for number of image samples for case and control conditions in testing sessions). To achieve balanced numbers of training samples for case and control conditions, resampling is performed based on different probability distributions for the two conditions. For control condition, downsampling is necessary to reduce the excessive number of images in order to speed up the training process. 10-second image samples are generated consecutively along the timeline with no overlapping. Then for each session, 10,000 images are selected based on uniform distribution to ensure randomness. On the other hand, oversampling is needed for case condition to obtain comparable quantity of case images to those of control condition. Based on an intuitive heuristic that samples closer to maxima of ST change hold more information about case condition, sample selection based on Gaussian distribution is proposed to augment the probability of image samples in the training set with center time points near maxima from both directions in timeline. For each event episode in the case condition, the distribution mean is set at the corresponding maxima of ST change, and standard deviation is set to be 10 seconds. This configuration ensures a large probability close to maxima of ST change, and 99% of image samples selected for training fall within 30-second radius of the maxima. With center time points selected based on the probability model, 10-second image samples are generated by
combining a 5-second ECG before and after the centers. For each session, an equal number of samples is selected from each event episode, and is set to be number of samples in control (10,000) divided by the total number of episodes in that session. Samples with partial information outside episode boundaries are excluded from case condition. In total, there are 174,039 case samples and 200,000 control samples generated from the 20 training sessions. For testing sessions, testing image samples from both case and control conditions are simply selected consecutively in the timeline to preserve real-time fashion.

Figure 1. Exemplar image samples from control and case conditions. (a) 10-second image samples with no significant ST changes; (b) 10-second image samples with significant ST changes.

2.3 CNN model from transfer learning

The proposed CNN model takes advantage of pretrained Google Inception V3 model\(^1\), which is trained upon millions of images and hundreds of classes from ImageNet\(^1\), through transfer learning scheme. The idea is to retrain the existing model, which already captures rich primitive features that are common to different image applications, to classify ECG images in the present study. Figure 2 presents the schema diagram for retraining CNN model from Inception V3 via transfer learning. Layers in Inception V3 model can be separated into two parts, the feature selection part and the classification part. The feature selection part includes convolution layers, max pooling layers, dropout layers, etc., while the classification part includes fully connected layer and Softmax layer. Within transfer learning paradigm, the feature selection part of Inception V3 is kept, and transferred features are generated by passing training samples in the present study through all layers in the feature selection part. Then final layers, such as fully connected layer and Softmax layer, are trained and appended based on the class labels (i.e., case and control) in the present study to yield final classification results. The final layers are trained and updated through 50 epochs. With the retrained CNN model, classification is performed for each image sample in the testing sessions.

Figure 2. Schema diagram for retrained CNN model via transfer learning
2.4 Performance evaluation

To investigate the performance of proposed CNN model, common metrics for accessing binary classifiers, including the receiver operating characteristic (ROC) curve and the area under the curve (AUC), are adopted to evaluate the model performance under various discrimination thresholds. The AUCs from all testing sessions are then tested against the random guess level (50% for binary classification) using one-sample Student’s t-test. For each testing session, an optimal cutoff threshold can be selected based on Youden’s index17:

\[ J = \frac{TP}{TP + FN} + \frac{TN}{TN + FN} - 1 \]

where J denotes Youden’s index, and TP, FP, TN and FN are counted by treating case condition as positive class. Optimal cutoff thresholds are selected where Youden’s indices reach maxima, and additional metrics, including sensitivity, specificity and F1 score, are calculated at the optimal cutoff thresholds to further evaluate the classification performance.

3. Results

The retrained CNN model from Inception V3 via transfer learning uses the first 20 sessions in the LTST database (174,039 case samples and 200,000 samples) as training, and it is tested on the following 15 sessions in the database. Figure 3 presents the ROC curves achieved by the proposed model for all 15 testing sessions. The black dashed line denotes guess level, and each blue curve denotes ROC curve of one testing session. The x axis represents 1-specificity, while the y axis represents sensitivity. It shows the proposed model is able to detect ST changes from control condition above the chance level for all sessions, while some variation across different sessions in performance also presents.

![Figure 3. ROC curves of individual testing sessions. Black dashed line indicates random guess level.](image)

Table 1 provides quantitative performance evaluation with various performance metrics from all testing sessions achieved by the proposed model. Each row presents performance information from one testing session, with last row presenting the mean and standard deviation derived from all 15 sessions. Specifically, the first column shows sessions in the LTST database selected as testing sessions. The second column (# case/control) presents number of
image samples from case and control conditions for each session. It reveals highly unbalance samples between both conditions. The third to last columns present AUC in correspondence to each ROC curve in Figure 3, and sensitivity (Sen@opt), specificity (Spec@opt) and F1 score (F1@opt) at selected optimal cutoff thresholds. The AUC demonstrates above chance performance for all testing sessions, with 9 out of 15 testing sessions reaching AUC value above 90%. In average, the proposed CNN model achieves mean AUC of 89.6±9.3% from the 15 independent testing sessions. In addition, the Student’s t-test suggests this performance is significantly higher than the random guess level (p<0.001). With selected optimal cutoff threshold, the proposed model achieves mean sensitivity of 84.4±13.9%, mean specificity of 84.9±8.3% and mean F1 score of 89.2±5.5%, respectively.

Table 1. List of performance information for each testing session.

<table>
<thead>
<tr>
<th>Session</th>
<th># case/control</th>
<th>AUC (%)</th>
<th>Sen@opt (%)</th>
<th>Spec@opt (%)</th>
<th>F1@opt (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>s20341</td>
<td>480/15702</td>
<td>99.1</td>
<td>96.9</td>
<td>96.1</td>
<td>96.8</td>
</tr>
<tr>
<td>s20361</td>
<td>1426/15700</td>
<td>98.6</td>
<td>95.4</td>
<td>92.6</td>
<td>93.7</td>
</tr>
<tr>
<td>s20331</td>
<td>346/17162</td>
<td>98.2</td>
<td>94.2</td>
<td>94.9</td>
<td>96.2</td>
</tr>
<tr>
<td>s20251</td>
<td>141/16968</td>
<td>97.6</td>
<td>97.9</td>
<td>91.5</td>
<td>94.9</td>
</tr>
<tr>
<td>s20351</td>
<td>523/15893</td>
<td>97.5</td>
<td>89.7</td>
<td>94.2</td>
<td>95.3</td>
</tr>
<tr>
<td>s20231</td>
<td>818/16032</td>
<td>95.8</td>
<td>95.2</td>
<td>87.9</td>
<td>91.0</td>
</tr>
<tr>
<td>s20211</td>
<td>391/16197</td>
<td>95.3</td>
<td>96.2</td>
<td>79.4</td>
<td>86.8</td>
</tr>
<tr>
<td>s20241</td>
<td>1161/15653</td>
<td>90.8</td>
<td>86.9</td>
<td>80.1</td>
<td>85.0</td>
</tr>
<tr>
<td>s20371</td>
<td>190/15924</td>
<td>90.7</td>
<td>87.9</td>
<td>78.9</td>
<td>87.2</td>
</tr>
<tr>
<td>s20261</td>
<td>1183/14899</td>
<td>87.7</td>
<td>77.7</td>
<td>84.8</td>
<td>87.3</td>
</tr>
<tr>
<td>s20281</td>
<td>262/16796</td>
<td>86.5</td>
<td>83.2</td>
<td>76.9</td>
<td>85.7</td>
</tr>
<tr>
<td>s20321</td>
<td>34/17647</td>
<td>81.0</td>
<td>79.4</td>
<td>78.0</td>
<td>87.4</td>
</tr>
<tr>
<td>s20311</td>
<td>1803/15345</td>
<td>80.2</td>
<td>67.4</td>
<td>85.5</td>
<td>85.7</td>
</tr>
<tr>
<td>s20291</td>
<td>722/15724</td>
<td>73.0</td>
<td>68.1</td>
<td>66.2</td>
<td>76.2</td>
</tr>
<tr>
<td>s20301</td>
<td>593/16601</td>
<td>71.7</td>
<td>49.6</td>
<td>86.0</td>
<td>89.0</td>
</tr>
<tr>
<td>Mean/STD</td>
<td>NA</td>
<td>89.6/9.3</td>
<td>84.4/13.9</td>
<td>84.9/8.3</td>
<td>89.2/5.5</td>
</tr>
</tbody>
</table>

4. Discussion

The present study introduced an image-based approach in combination with a deep learning technique to monitor ischemic ST change in ambulatory ECG. A CNN model has been trained through transfer learning scheme using 24-hour ambulatory ECG recording sessions in LTST database and tested on independent sessions in the database. The proposed CNN model is able to classify testing images in real-time fashion with an average AUC at 89.6%, which is significantly higher than guessing (p<0.001). At selected optimal cutoff thresholds, our model achieves an average sensitivity at 84.4% at 10-second sample level. This is on par with previously reported performance using the same annotation protocol with average sensitivity at 78.10%, 78.28% and 82.13%, but achieved at whole-episode level, i.e., assigning just one class label for the whole duration of ST episode. Meanwhile, our model is able to maintain a high specificity, demonstrated by the average specificity at 84.9% and F1 score at 89.2%.

To our knowledge, the present study is among one of the first studies leveraging deep learning technique for detecting ischemic ST change in ECG. The proposed approach delivers a simple training process in comparison to
previous ST detection algorithms\textsuperscript{1,4}, while achieves comparable if not higher performance. It bypasses the complicate rule settings from previously used decision tree, which might increase the chance of overfitting. The image-based sample selection resembles images presented in ECG monitors for human inspection. Such approach is inspired by our previous study that presented ECG signals as images and differentiated images of good ECG quality from those of poor quality\textsuperscript{9}. Both approaches are motivated by an insight that clinicians typically use visual pattern recognition to read ECG and identify pathological changes in ECG tracings. The combination of image-based sample selection and deep learning produces a model with both high sensitivity and specificity for monitoring significant ST change at 10-second sample level, which is crucial for early detection of myocardial ischemia. Accurate sample-level classification is also the foundation of successful detection of whole episode. Thus, our proposed approach also contributes to the effort of reducing false alarms that plague current ST monitor systems.

One limitation of present study resides in annotations for sudden-step ST change caused by axis shift and conduction change in LTST database, which are only provided with onset time. Except for the sharp change at the episode boundaries, samples associated with these events can be akin to those from ischemic ST change, and the duration varies from episode to episode. Although we have mitigated the issue by removing empirically selected 1-minute ECG centered at their onsets from control condition, residuals could still greatly limit the specificity. In addition, the large variation in number of sudden-step ST episodes across patient-independent testing sessions could shed light on the performance variation as presented in Figure 3. Thus, additional measures of event duration are needed to have these events truly accounted for during model training.

Despite the achieved performance, the present study is still preliminary and orientated as a proof of concept, with only 20 out of total 86 sessions in LTST database included in training set and 15 in testing. Future effort is needed to investigate model performance with respect to number of sessions included in the training set, as well as to validate the stability on a larger testing set. Furthermore, to expedite the training process, only 50 epochs of optimization is performed in current model, while performance improvement is expected with further optimization. Due to highly-unbalanced data in case and control, an oversampling approach based on Gaussian distribution is introduced in case condition. Our results demonstrate the efficacy of such approach in capturing representative samples in case condition. However, the selection of 10-second STD for the probability distribution is empirical. We also plan to systematically investigate the impact of its changes on the model performance in future work.

Acknowledgement

This work in partially funded by R01 NHLBI128679, R18H5022860, R01GM111378, UCSF Middle-Career scientist award, and UCSF Institute of Computational Health Sciences.

References


Application of Data Provenance in Healthcare Analytics Software:
Information Visualisation of User Activities

Shen Xu, PhD\textsuperscript{1}, Toby Rogers, BSc\textsuperscript{2}, Elliot Fairweather, PhD\textsuperscript{1}, Anthony Glenn, BA\textsuperscript{2}, James Curran\textsuperscript{2}, Vasa Curcin, PhD\textsuperscript{1},
\textsuperscript{1}King’s College London, London, United Kingdom; \textsuperscript{2}Imosphere Ltd, Nottingham, United Kingdom

Abstract

Data provenance is a technique that describes the history of digital objects. In health data settings, it can be used to deliver auditability and transparency, and to achieve trust in a software system. However, implementing data provenance in analytics software at an enterprise level presents a different set of challenges from the research environments where data provenance was originally devised. In this paper, the challenges of reporting provenance information to the user is presented. Provenance captured from analytics software can be large and complex and visualizing a series of tasks over a long period can be overwhelming even for a domain expert, requiring visual aggregation mechanisms that fit with complex human cognitive activities involved in the process. This research studied how provenance-based reporting can be integrated into a health data analytics software, using the example of Atmolytics visual reporting tool.

Introduction

Because of the increasing complexity of analytical and data tasks, the aim of analytics software is to devise and construct visual abstractions in addition to multifaceted information so to provide useable options. New data protection laws and regulations from the EU have demanded and stipulated that computer-assisted choices must be explainable\textsuperscript{1}. Furthermore, this requirement tackles the “black box” created around the reasoning behind important actions, increasing the risk of both human error and systemic errors, and engendering mistrust in analytics\textsuperscript{2}.

Data provenance is a technique that describes the history of digital objects, where they came from, how they came to be in their present state, who or what acted upon them. Provenance maintains the integrity of the digital objects, e.g. the results of a data analysis engender greater trust if their provenance shows how they were obtained. In health data settings, it can be used to deliver auditability and transparency, and to achieve trust in a software system. Specifically, in Learning Health Systems, it is applicable across a range of applications\textsuperscript{3}. Provenance data is commonly represented as directed acyclical graphs, with vertices being data entities, processes operating on those data and agents controlling those processes, and edges being causal relationships between the different types of vertices (e.g. data produced by a task, task utilising some data, etc).

Despite the interest in the field, there is still much work that needs to be done to make provenance practical. One important step is to obtain a meaningful answer to the query, “how was this object produced?” which may be hindered by large volumes of collected provenance data. On an enterprise level of system development, the query needs to be answered by responding to a rational justification as well as regarding a semantic viewpoint\textsuperscript{4}. Rational justifications in this sense are not able to be automatically derived from software processes, indeed they demand input from the user; however, the user’s rationale can be assisted or enabled through provenance data capturing. The exemplar software used in this paper, Atmolytics, is a healthcare focused analytics system, interacting with patient datasets to provide answers to specific user questions. When loaded with social care data and care cost information, Atmolytics can be used by a care services director who would like to know if there are service users whose needs can be met in a more cost-effective way.

It is clear that the research community is aware of the necessity of supporting provenance; according to the taxonomy of provenance technology\textsuperscript{5}, many researchers have developed tools and systems that help provenance dissemination, analysing records both workflows \textsuperscript{6–8} and reasoning processes\textsuperscript{9–14}. Tools designed to support provenance across domains, particularly using an automated capture method, typically capture event-based provenance (e.g. clicks, drags, and key-presses) at a very granular level, consequently increasing computational cost and potentially overwhelming end-users.
Our research focuses on insight provenance, which shares the view of Gotz and Zhou\textsuperscript{13}, which described how the HAVEST system captures the history of actions during financial analysis activities by identifying a set of semantic units of user activity. This research further provides links between semantic actions and data entities, then represents them in an organized timeline to facilitate consumption of provenance data by end-users. Such difficulties and challenges are covered in the second section of this paper, with the focus on existing designs and challenges within the user interface (UI) design. The Atmolytics analytics software use case shall be outlined in the third section. After the user test results have been shown and assessed, a conclusion shall be drawn.

**Challenges of visualising provenance**

Provenance of digital scientific objects is metadata that can be used to determine attribution, to establish causal relationships between objects, to find common tasks and parameters that produced similar results, as well as for establishing a comprehensive audit trail to assist a researcher wanting to reuse a particular data set\textsuperscript{15}.

However, considerable effort is required to ensure the usability of provenance. To illustrate this, consider a typical provenance question: “What is the means by which the object in question was created?” Answering this question may be hindered by large amounts of provenance data collected. According to Macko et al. (2013), introducing local clustering into provenance graphs enables the identification of a significant semantic task through aggregation. This may be realised through the utilisation of online and offline metrics and a detection algorithm that may generate significant or informative results. As an example, when commencing with a seed node, S, only those nodes from S’s ancestry corresponding to semantically significant actions are present in the cluster – those that are produced by S which may also be built to assess and evaluate among and between numerous thresholds. Data provenance generation was introduced by Peng Chen et al. (2014) through the use of logical time, though this was achieved without provenance volume control generated at the source\textsuperscript{16}.

Finely grained provenance capture of digital scientific data could yield a graph that the user may find overwhelming in detail. This huge amount of data can be processed in a number of ways, e.g. by caching certain content or by progressing and furthering provenance perspectives\textsuperscript{17,18}. Provenance capture may also be throttled so that the number of provenance data generated at the source can be controlled\textsuperscript{19}. Regardless of these techniques, visualisation techniques are crucial in interpreting provenance data\textsuperscript{20}.

An absence or insufficient system engineering perspective must be accounted for, and this is the assumption of practicable concerns and considerations regarding provenance application. Though it was devised with the intention of an engineering approach, the temporal representation herein used the same logical time. The main difficulties and challenges of implementing provenance in visual data analytics are detailed below.

- **Design patterns (multiple).** Before a particular design pattern can be followed, a system communication backbone must first be utilised (an example of this is an enterprise service bus). Web applications and web services may also possess certain design patterns (for instance, the MVC/PAC). Indeed, a particular design pattern may also apply to database access (such as an entity framework). Several different patterns for design demand varied and differentiated granularity levels. A multistate approach is most popular for web services and web applications, and, in addition, the MVC (model-view-control) and presentation-attraction-control (referred to hereafter as PAC) are generally implemented in addition to the multi-layered architecture\textsuperscript{21}. Service-oriented architecture (SOA) provenance data was studied by Tsai et al. (2007), taking the data from a service framework viewpoint; one must consider web service/application data provenance on top of SOA. Indeed, enterprise software can process numerous data at the source, at the destination of the data and in intermediary nodes. Certain information does not need to be represented to the software user, for example the data flow between model view controller.

- **Flexible combined service.** So that several different aims and intentions can be realised or met, services may be used in parallel or combined in some other manner.

- **Security.** To ensure data consistency among the collated and recorded set, sensitive data can be hidden or obfuscated.

- **Classification of dynamic data.** Not all data need to be traced within analytics software. Therefore, only those significant or important data need to be tracked. As a result, the classification of data by their criticality regarding provenance is implied as a result.
**Data literacy.** The capacity to comprehend the meaning and importance of data is referred to as data literacy. This incorporates the means by which charts and graphs are read and understood, the right extrapolations from the available data and the acknowledgement of data use regarding inappropriate or incorrect manners. Software users, even those using advanced analytics packages, have a wide range of data literacy.

So that such provenance may be differentiated from that of a scientific data provenance research study, comparison of traditional scientific and analytic software data provenance is shown in Table 1 below.

**Table 1. Comparison of analytics software data provenance and scientific data provenance.**

<table>
<thead>
<tr>
<th>Features</th>
<th>Scientific data provenance</th>
<th>Analytics software data provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design patterns (multiple)</td>
<td>N/A</td>
<td>Many design patterns can be involved, for instance, SOA and/or MVC which is usual for web applications driven by data</td>
</tr>
<tr>
<td>Services</td>
<td>N/A</td>
<td>Combination of methods by the user for adherence to different aims and goals</td>
</tr>
<tr>
<td>Security</td>
<td>Traditionally utilised security mechanisms such as covert channels, digital signatures, encryption, kernel authentication for security and authorisation</td>
<td>Same problems tackled, though within an SOA system context, particularly regarding data routing with content-based data</td>
</tr>
<tr>
<td>Data classification</td>
<td>Generally static</td>
<td>Potentially dynamic</td>
</tr>
<tr>
<td>Data literacy</td>
<td>High exposure level</td>
<td>Assorted</td>
</tr>
</tbody>
</table>

**User Activity Driven Solutions for Analytics Software**

Utilisation of provenance in regard to a particular domain is dependent on granularity level. Coarse- and fine-grained as concepts are relative to data being observed. For example, when applying data provenance to a relational database, tuple-level is referred to as fine-grained while relation-level is referred to as coarse-grained data.

Meanwhile, two granularities can be determined regarding SOA provenance:

- **Fine-grained provenance:** This refers to intra-system data movement tracing, such as where the data originated and where it is headed, the time at which the data was created, the rationale concerning the dataset and the data's termination.
- **Coarse-grained provenance:** This alludes to workflow processing and the data generated by it.

Within the context of Web application which is on top of SOA, whether the user is more concerned with the system processes’ ancestry or whether or not their colleague has arrived at their conclusion is the main issue in question. According to Roberts et al. (2014), provenance may be categorised into the following strata: *data level*, which concerns inter database movement; *the analysis level*, which concerns intra-systematic interactions; and finally, the *reasoning level*, which concerns the decision-making process, the reasoning and the thinking that comprise the function of the analytic stages. The provenance question, per these stages and this approach, may be translated into two different user stories. Further, it may be considered according to these three aforementioned levels. This can be seen in Table 2.
From Table 2, we can see that the same provenance question can be answered from different perspectives with different foci. Analytic provenance (for example, select and click) was included in our initial design of provenance template. Nevertheless, such interactions could be aggregated into abstract process through the relation of intangible systematic element, thus a bridge of system activities to reasoning steps can be ‘constructed’. Additionally, other studies have presented a machine-learning method that incorporates inference of reasoning provenance taken from log data of user interactions\textsuperscript{26}. The authors of this study believe that, as it is such a challenge to capture reasoning provenance data since it is fit within “the mind” of the analyst in question and more tacit than overt, though conversely the inter-system analyst interactions (analytic provenance) may actually be captured with relative ease, thereby allowing for action-expressed intentions. The endeavours of this research study concerned low-level classification, like scrolling, to reasoning stages regarding instance seeking. The significance of reasoning processes in analytics software is thus addressed.

According to the requirements of the SOA, this research extends the analytics software requirements with suggested mechanisms. This, it is hoped, shall bring about a more practicable use of and regarding data provenance. The requirements and mechanisms of this are summarised in Table 3. The mechanisms proposed, as per a system engineering perspective, adhere to a provenance template methodology as devised by another study, and it brings about a system architecture for provenance captured and displaying for web applications\textsuperscript{27}. Herein the representation of data provenance for software users constitutes the primary research focus.

Table 3. Suggested mechanisms.

<table>
<thead>
<tr>
<th>Item</th>
<th>Feature</th>
<th>Type</th>
<th>Analytics software requirements</th>
<th>Proposed mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Design patterns (multiple)</td>
<td>MVC/Entity Framework/SOA, and others</td>
<td>Should permit flexible capturing means</td>
<td>Provenance Abstracted Activity/template (W3C standard complied \textsuperscript{28,29}) + Selected feature</td>
</tr>
<tr>
<td>B</td>
<td>Services (Flexible combination)</td>
<td>Potential recombining of methods to meet different ends</td>
<td>Flexible capturing mechanism required</td>
<td>Provenance grafting/template method</td>
</tr>
<tr>
<td>C</td>
<td>Security</td>
<td>Disguising data</td>
<td>Should allow encryption on data sent across boundaries.</td>
<td>Pre-processing Encrypt and Decrypt/data</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>Authentication mechanism needed to guarantee data is taken from correct sources</td>
<td>Data/template accessibility capturing provenance</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>Different access needs to be provided to different users</td>
<td>Provenance</td>
<td></td>
</tr>
</tbody>
</table>
so to guarantee data security. Capturing/template agent node security

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>Data volume and assorted data literacy</td>
<td>Confidence/trust judgement concerning reasoning level</td>
<td>Representation allowed by adhering to the logical steps taken by the user</td>
</tr>
<tr>
<td>G</td>
<td>Judgement of confidence/trust above that of analysis level</td>
<td>Support should be provided for logical steps</td>
<td>Feature list (3) + Highlighted difference</td>
</tr>
<tr>
<td>H</td>
<td>Judgement of confidence/trust above that of data level</td>
<td>Support should be provided for data sources</td>
<td>Summary of Linked System Activity</td>
</tr>
</tbody>
</table>

Partitioning of Provenance Data Graphs

There are two reasons why an annotated provenance graph is not the best visual solution for representing provenance data. Firstly, with potentially thousands of attributes and nodes in a graph, scoping the relevant items becomes difficult. Second, placement of non-structural and structural information within a single status display yields unwieldy results. As a result of this, we have introduced a graph partitioning mechanism based on temporal representation and user activities, with the aim of producing confidence in a given data product within an analytics solution. Further relevant information concerning the system-based activities and related automated processes is also included.

A provenance graph contains three different types of nodes: agent, activity and entity nodes. Our partitioning is total and creates a series of non-overlapping node subsets according to the logical temporal stamp and the targeted identifier. A subset of a provenance graph is therefore, more exactly, seen to be \( G=(V, E) \). Here, the set of agent nodes is denoted by \( V^a \), while the set of entity nodes are denoted using \( V^e \), while the activity node set is denoted by \( V^c \); \( V^a \cup V^e \cup V^c = V \) is used to denote the total set of nodes wherein \( E \) denotes all edges, which can be seen to be denoted per: \( G=(V, E) \) provenance graph, \( V \) is then partitioned into several subsets, per \( V_1, V_2 \) etc. Thus:

1. \( V_1, V_2, \ldots, V_k \in V \) and \( \bigcup_{i=1}^{k} V_i = U \)
2. \( \forall i \neq j \) and \( 1 \leq i, j \leq k \), \( V_i \cap V_j = \emptyset \)
3. \( \forall a, b \in V^c \subseteq V_i \), we must have \( \text{LTS}(a) = \text{LTS}(b) \)

LTS is used to denote the logical temporal stamp. Here a function assumes an activity node to be an output and thus generated an output stamp. Additionally, each node is annotated with the stamp according to its adherence to certain relationships. One-day-long provenance graphs are made from the present testing stamp (which is one day long).

![Figure 1](image.png)

Figure 1. Activity driven representation of data provenance. Every one of the subsets is arranged into an ordered list as such \( \{V_1, V_2, \ldots, V_k\} \). An “appears prior to” relationship was suggested by Chen et al.\(^6\). The approach used herein is to arrange such subsets into a timeline arrangement per (logical stamp) real-time ordering. Hence throughout the capturing process, the subsets will be ordered sequentially. This implicit order is then undertaken as the user interacts with the web app. This is in contrast to some other studies in the provenance field, many researchers that ignore the temporal ordering and instead use significant computational activities to re-engineer it from the bottom-up. The activity design is shown in Figure 1, to be read from left to right, showing our design of activity-based temporal representation.
User tests and analytics software case study

In the Atmolytics architecture, an enterprise service bus is used to receive data tasks before they are distributed over several farms for processing, enabling multiple shared databases to be used to complete a single task. The means by which a provenance infrastructure can be embedded into the Atmolytics system architecture is depicted in Figure 2, based on programmatic calls within Atmolytics invoking a RESTful API upon the provenance server. The provenance services correspond to standard actions in the system and are implemented using abstract provenance templates\textsuperscript{11, 12} which get instantiated during API service calls with concrete data and persisted into the provenance data store\textsuperscript{27} (B in Figure 2). Provenance capturing (A in Figure 2) is triggered by a controller in Atmolytics – a Targeted Activity – after which the object continues to the method used for data processing. Content may be encrypted prior to its being sent to data storage regarding every individual data slot. This enables our proposed mechanism for security requirements in table 3, Item C. The use of provenance templates enables our proposed mechanism for design patterns and security in table 3, Items A, D, E. After a new graph is created (C in Figure 2), it is typically linked into the existing graph by means of grafting the new nodes onto the existing structure. This supports the loosely-coupled services feature in table 3, Item B.

Figure 2. Simplified representation of web application processes and provenance capturing (adapted from Xu et al 2016\textsuperscript{30})

A: Provenance capture triggered from a process in Atmolytics;
B: Abstract provenance template gets instantiated with the details provided by the process\textsuperscript{28, 29};
C: New provenance data gets grafted onto existing data in the provenance data storage\textsuperscript{11, 12}. 

268
The initial round (comprised of five users) of the user advisory group is completed, according to present progress. This intends to ascertain a collated sentiment or opinion as well as design feedback above that of the implementation of provenance and thus assist the system regarding its trust/confidence in data products as well as increase transparency. Every one of the tests is approximately two hours in length and the analytics software demonstrative exemplar is covered in each of the sessions. Feedback regarding the designs is provided as and when the user interacts with the mocked-up interface

Figure 3. User experience review mock-up of user interface (https://invis.io/NQDCUN6DS shows the 2nd round user interface). F1 is the logical temporal stamp; F2 is the activities’ justification; G3 is the prechosen feature list; G4 underlines the changes; and H is the linked system activity summary, refer to table 3.

Looking at Figure 3, F1 shows the revised design of temporal representation, because during our review, some users found the aggregated nodes showing number of activities confusing. Compared to Figure 1, all nodes are of equal size
and allocate them distinctive, representing the requirement F in Table 3. Annotation/Justification (F2) allows user free
text to be shown alongside the provenance information. Data exploration is a form of unstructured problem solving,
so activities should be justified in case of reviewing. Together with F1, the reasoning steps and justification provide a
higher-level auditing of data exploration.

G3 denotes the feature space of each activity, describing how the data is affected by the user-selected features. According to the user feedback, we have highlighted the changes between activities (G4). H represents system activities involving system processes linked to the changes of cohort, for example updates, changed base cohort size etc.

The users were also asked about the sentiments and opinions concerning the implementation of our provenance infrastructure. Their opinions are summarised in Table 4.

**Table 4.** Summary of user feedback. (5 Strongly agree, 4 Agree, 3 Neutral, 2 Disagree, 1 Strongly disagree).

<table>
<thead>
<tr>
<th>Questions</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 The provenance information displayed helps you understand how the result is produced.</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2 The provenance information displayed provides transparency to the processes of producing outputs.</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>3 The provenance information displayed improves the trust/confidence of outputs.</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4 The provenance feature captures the decisions involved in producing an output.</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5 The provenance feature is useful to your work.</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

The initial feedback from these users indicates that our approach aligns well with users’ own cognitive model. For example, all users during the test session recognize and understand well of displayed activities. One user reported that “If accompanying annotation/comments it can make clear the rationale”. In addition, all users welcomed a future review of the design to see the development of provenance feature.

Notably, user B disagreed with the confidence/trust statement. “At a basic level yes but the problem as discussed is the lack of transparency back to the original distributed data sources themselves. The outputs are dependent on that data so ideally you need to have visibility at least of the import/mapping process to those data sources to have trust in the report outputs”. This relates to the stage in the data lifecycle at which we start capturing the provenance. In our current implementation, this begins once the data sets are in the Atmolytics system. However, we are currently exploring the implementation of provenance during the Extract-Transform-Load stage during which original data is being transformed and loaded into Atmolytics, and manually annotating the loaded data with relevant ethics and governance information.

**Conclusion**

Provenance is a critical requirement for analytic applications, but the methodologies of many existing implementations, typically manual browsing of recorded provenance or pre-defined queries of provenance data, have fundamental limitations. This study aims to establish a flexible method of data provenance capture and visualization within analytics software, using the Atmolytics health data analytics software as an exemplar. As part of the study, the objective of this paper is to develop a new approach based on semantic actions that combines the benefits of both approaches while avoiding their deficiencies.

The research concentrates on the various analytic processes’ reasoning steps, with the consequential design being more appropriate to a broader provenance community, and potentially driving wider appreciation and adoption of provenance. A positive response to the activity-driven temporal representation of data provenance is confirmed in our initial evaluation, and user studies are continuing.

Several challenges remain to be addressed in future work. In particular, our approach represents provenance artefacts generated by system-level processes in the same way as user-generated events, but the abstraction required to do this may cause a loss of fidelity for some such processes; further development could produce an activity map that has capacity to increase the granularity of the activity to address this. Finally, we must perform more comprehensive system performance evaluation to fully evaluate our system development in addition to user studies.
References

1. Association for Computing Machinery US Public Policy Council (USACM). Statement on algorithmic
2. Broadhurst K, White S, Fish S, Munro E, Fletcher K, Lincoln H. Ten pitfalls and how to avoid them: What
research tells us. 2010;(September):1–42. Available from: papers2://publication/uuid/3ABE0587-8B27-
4BEB-9A22-F3B4B34D4C85
3. Curcin V. Embedding data provenance into the Learning Health System to facilitate reproducible research.
Learn Heal Syst [Internet]. 2017 Apr 1;1(2):e10019–n/a. Available from: http://dx.doi.org/10.1002/lrh2.10019
for Integrating Syntax, Semantics
and Pragmatics for Computer-aided Professional Practice: With Application of Costing in Construction
http://www.scopus.com/inward/record.url?eid=2-s2.0-57149126952&partnerID=Z0tx3y1
6. Parker SG, Johnson CR. SCiRun : A Scientific Programming Environment for Computational Steering. In:
data - SIGMOD '08 [Internet]. 2008;1345. Available from:
http://www.scopus.com/inward/record.url?eid=2-s2.0-57149126952&partnerID=Z0tx3y1
Multiple-View Visualizations. In: IEEE Conference on Visualization [Internet]. Minneapolis, MN: IEEE;
11. i2 i. Analyst’s Notebook [Internet]. IBM; Available from: https://www.ibm.com/uk-
-en/marketplace/analysts-notebook
26.
Intell Lect Notes Bioinformatics) [Internet]. 2004;3298:92–106. Available from:
http://twiki.ipaw.info/bin/view/Challenge/MINDSWAP
Collection. In: Moreau L, Foster I, editors. Provenance and Annotation of Data: International Provenance
and Annotation Workshop, IPAW 2006, Chicago, IL, USA, May 3-5, 2006, Revised Selected Papers
[Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 171–83. Available from:
https://doi.org/10.1007/11890850_18
20. Silva CT, Freire J, Callahan SP. Provenance for Visualizations: Reproducibility and Beyond. Vol. 9,
2014;(May).
and Research Faculty. 2011;
A Crowdsourcing Framework for Medical Data Sets

Cheng Ye, Joseph Coco, Anna Epishova, Chen Hajaj, Henry Bogardus, Laurie Novak, Joshua Denny, Yevgeniy Vorobeychik, Thomas Lasko, Bradley Malin, Daniel Fabbri
Vanderbilt University, Nashville, TN, USA

Abstract

Crowdsourcing services like Amazon Mechanical Turk allow researchers to ask questions to crowds of workers and quickly receive high quality labeled responses. However, crowds drawn from the general public are not suitable for labeling sensitive and complex data sets, such as medical records, due to various concerns. Major challenges in building and deploying a crowdsourcing system for medical data include, but are not limited to: managing access rights to sensitive data and ensuring data privacy controls are enforced; identifying workers with the necessary expertise to analyze complex information; and efficiently retrieving relevant information in massive data sets. In this paper, we introduce a crowdsourcing framework to support the annotation of medical data sets. We further demonstrate a workflow for crowdsourcing clinical chart reviews including (1) the design and decomposition of research questions; (2) the architecture for storing and displaying sensitive data; and (3) the development of tools to support crowd workers in quickly analyzing information from complex data sets.

1 Introduction

Crowdsourcing has gained notoriety as services like Amazon Mechanical Turk (AMT) have enabled researchers to ask questions to crowds of workers and quickly receive labeled responses. These human labeled data sets are increasingly important for training supervised machine models, as labels do not exist for many important research questions and cannot be produced with automated methods. Unfortunately, crowds composed of individuals from the general public are inappropriate for numerous types of data sets that require crowdsourcing, such as clinical data, due to legislation (e.g., the Health Insurance Portability and Accountability Act of 1996) and organizational policies. In particular, privacy concerns prevent arbitrary users from accessing these data. Moreover, the subject matter being analyzed requires highly specialized training and expertise to accurately produce a label, which is often not available in a public crowd.

This paper outlines a crowdsourcing framework for medical data sets and one current deployment of the system. There are many components necessary for building such an environment to allow for scalable human computation on medical data sets. Broadly, the main components of the system include: (i) a crowdsourcing system that can be deployed within an organization that has the ability to specify workers’ attributes, roles and access controls, (ii) de-identification routines to perturb identifiers and meet ethical and legal requirements, (iii) graphical user interfaces to display sensitive data, (iv) and machine learning tools to assist workers to produce labels quickly. Moreover, beyond the technical components, this paper describes organizational processes that are needed to train researchers about crowdsourcing so they can construct well-defined questions for the crowd, and approaches to recruit skilled workers.

To demonstrate the challenges and complexities of developing and deploying a crowdsourcing system for sensitive data, this paper focuses on the important use case of clinical chart reviews. Chart reviews are a common component of medical research in which medical students, staff or nurses comb through semi-structured electronic medical records (EMR) systems (which are designed for clinical treatment rather than research) for specific data. Unfortunately, scrolling through vast amounts of clinical text to produce labels is time-consuming and expensive. For example, at Vanderbilt University Medical Center, it currently costs $109 per hour for a service which pays a nurse to review patient charts and produce labels, where a large part of this fee goes to project management and other overhead. While some researchers have employed software scripts to infer labels from text data automatically, the messiness and complexity of EMR systems’ semi-structured data make verifying the accuracy of the results difficult. More problematic is that the clinical text is filled with misspellings, medical acronyms, and abbreviations which make disambiguation difficult with natural language processing techniques.

One major challenge for crowdsourcing workers is uncovering relevant information quickly from complex data sets. For example, in healthcare, patient charts are managed as a collection hundreds, if not thousands, of clinical docu-
ments, each of which may include tens of pages of information. Finding the specific paragraph related to a patient’s diabetes care history or cancer medication adherence is nontrivial. While keyword search can help find some content, variations in terminology and other clinical semantics make finding all relevant data challenging. Moreover, identifying all relevant text in a single note related to, say, seizures remains time-consuming and requires extensive skimming. For these reasons, the crowdsourcing framework requires additional tools to assist workers in finding relevant content quickly, such as text highlighting and data visualization for summarization.

An effective crowdsourcing system for medical data sets can change how medical research is done and allow researchers to solve important problems. In our experience, the chart review process is often a key rate limiting step for modern studies; crowdsourcing has the ability to substantially lower the time to complete clinical studies. Additionally, the resulting labels are invaluable resources for supervised machine learning researchers that otherwise would be limited by smaller training data sets.

Consequently, as displayed in Figure 1, our goal is to build a lightweight, customizable pipeline that significantly reduces the cost and time to complete medical research while increasing reproducibility and accuracy, and maintaining privacy and security standards.

2 CROWDSOURCING FRAMEWORK

This section introduces the main components of the crowdsourcing system, defines the workflow in which researchers develop crowdsourcing questions, and describes the process workers follow to produce labels.

1. Questions Design: After recruiting and identifying a researcher with an amenable research project, we conducted a design workshop. The workshop included the researcher, medical personnel, computer science researchers and anthropologists. The workshop began by introducing the researcher to crowdsourcing preliminaries and non-healthcare crowdsourcing examples. Next, the team worked to clarify and decompose the research objective into atomic questions.
by refining the structure of the crowdsourcing project as in Figure 2. We discussed data needs (e.g., all notes or specific note types), question format (e.g., boolean, multiple choice or text snipping), scope of tasks (e.g., multiple questions per patient or a single one), and worker skills requirements for the tasks.

Crowdsourcing questions were constructed as narrowly as possible. For example:

(a) Does a clinical note document patient conversations regarding diabetic diet alternatives? (Y/N)

(b) Which of the following dietary alternatives were discussed with the patient? (Healthy oil choices, Sugar free sweets, Unsweetened tea, None)

(c) Of the patient’s current diet choices listed in this note, rank them in terms of most problematic to their long-term health: (Soda, French fries, Dark chocolate, Broccoli)

In addition to true/false questions, multiple choice questions and ranking questions, researchers also asked that workers snip (or extract) text from notes that support the answer. We found these snippets were extremely helpful when experts were needed to adjudicate disagreements.

2. Data Extraction and De-Identification: APIs are needed to extract data from the underlying data store and load them for analysis. In an academic setting, these APIs are configured to take an Institutional Review Board (IRB) number and return the set of medical record numbers in the study. For each medical record number (MRN), the associated charts are pulled and loaded into the system. Moreover, upon querying the charts, the APIs apply open-source, de-identification tools (e.g., the MITRE Identification Scrubber Toolkit\textsuperscript{13}), to remove or scrub HIPAA-designated identifiers, such as patient name and residential addresses.

3. Crowdsourcing System: At its core, a crowdsourcing system matches workers to questions and stores the answers (or labels) in a database. Instead of developing a crowdsourcing system from scratch, we leveraged the open-source Pybossa\textsuperscript{14} system. Pybossa provides many basic crowdsourcing features, such as: loading and styling questions (known as a presenter), registering workers, assigning workers to tasks, collecting answers, timing tasks and extracting aggregate statistics and labels.

Unfortunately, the default version of Pybossa lacks many of the privacy controls that are needed to manage sensitive data. Therefore, fine-grained access controls with two-factor authentication were added to limit access for each worker. Moreover, worker attributes (or properties) were added to the underlying worker data models so each worker could be categorized by his or her skill level and specialty. These attributes allow for fine-grained question assignment and weighting.
The resulting crowdsourcing system was deployed on an internal server within the Vanderbilt University Medical Center firewall. The site was not open to the public. All worker registration, task assignment, question answering and data extraction were managed through a web interface over HTTPS and the activity is logged.

The Pybossa system allows researchers to customize how questions are ‘presented’ to workers via basic HTML and JavaScript coding. These presenters are simple templated HTML forms that read from an API and populate question text and candidate answers.

4. **Customizable Display Interface:** We specifically decided to separate the crowdsourcing system from the data display system. This abstraction separated the logic of presenting questions to workers from the task of effectively displaying data for the specific research project. Instead we used HTML IFrames as a means for Pybossa to point to data for analysis. These IFrames can load content from a given URL and provides the developer control over the data input, method of display and tools used to parse the data. The IFrame URL is another parameter specified when configuring a Pybossa project.

The IFrame design has proven to be extremely versatile as we have completed projects displaying different data types including clinical text and medical images.

5. **Helper Libraries:** Perhaps the most important component of the system is a set of helper libraries that assist workers to produce labels. Example helper libraries include text highlighting tools, text search and document ranking.

6. **Worker Recruitment:** When working with sensitive data, only certain individuals have the necessary credentials to access the data. For instance, in healthcare, only hospital employees (which includes faculty, staff, and trainees) can access medical records. While the pool of workers is limited (in contrast to the aforementioned public crowd on Amazon), there are often groups of highly motivated workers, such as medical students, who are willing to work given incentives.

Our worker pool consists of mostly medical students and nursing students, with a small number of faculty. These workers were recruited through Grand Round presentations and IRB-approved email communications. For a worker to participate, he or she signed a data use agreement and, in some cases, was added as key personnel to the researcher’s IRB. We also recorded skill-level of each worker (e.g., medical student, intern, resident, fellow, attending, and nurse) and their specialization (if any), as these answers can impact which questions they are qualified to answer.

7. **Supervised Machine Learning:** After the crowdsourcing task is complete, researchers can use the labels to train supervised machine learning models. In healthcare, popular prediction systems include clinical decision support and order recommendation, among many others.

3 **EXAMPLE USE CASE**

In this section, we present an example use case of the crowdsourcing framework.

Suppose a researcher needs to label notes from a diabetes cohort (e.g., patients with ICD code 250.*). For each note, a worker selects one of the following labels: not relevant, relevant, or partially relevant to diabetes care. Moreover, for a note with a relevant or partially relevant label, the researcher also wants to extract supportive snippets from the note.

![Figure 3: Example of Pybossa presenter with a text search engine.](image)

After clarifying the research question, task scope, task corpus and worker action, a presenter is designed and loaded into Pybossa. Similarly, notes are extracted from the EMR system, de-identified and loaded into a chart review data system.
Workers are recruited and assigned to the project. A pre-test determines if each candidate worker has sufficient knowledge about diabetes to participate. Only candidates who pass the test are admitted into the worker pool and assigned tasks.

Admitted workers then begin reading notes assigned to them by Pybossa and producing labels. One note is shown to each worker at a time. The worker reads the content of the note, chooses a label, and selects relevant snippets from the note. This process continues until all notes are labeled. Depending on the coverage requirements, multiple workers might answer the same question.

To reduce the cognitive load of workers, the Pybossa presenter (as shown in Figure 3) only consists of the question, answer options and the input box of the search engine. Moreover, we only select necessary helper tools to assist workers find relevant information.

As shown in Figure 4, the interface highlights “diabetes” and semantically similar terms of “diabetes”, such as “hyperlipidemia” and “obesity.” On the left side of the note, a heat map displays the number of terms related to diabetes in each section of the note. When multiple notes are displayed to a worker in a single task, an EMR search engine automatically finds notes that contain diabetes or similar terms, and ranks the notes using an information retrieval metric.

![Evaluation Note 9](image)

Figure 4: An example helper library: highlighting similar words in a note

After all tasks are completed, the researcher receives the labels and snippets. The researcher then utilizes the data in a supervised machine learning task, such as document classification.

4 HELPER LIBRARIES

Clinical notes contain vast amounts of unstructured information describing a patient’s past medical history and diagnoses. Ideally, crowd workers would be able to quickly parse through these data to answer their crowdsourcing
questions. Unfortunately, searching for relevant data remains difficult. While workers deploy basic routines, such as starting with the most recent note or a specific note type (e.g., discharge summary), these basic approaches can miss relevant data.

The objective of the helper libraries is to make common crowdsourcing operations more efficient. In the case of chart reviews, this means more efficiently finding relevant clinical text. To that end, we have deployed a set of search tools to find and rank documents for review.

We have implemented and tested three types of search systems:

- **Keyword search**: Return documents that contain the search term. Rank documents by the frequency of the search term in the document.
- **Expanded keyword search (Figure 5)**: Given a search term, expand the search to include terms that are semantically similar to the term (semantic similarity is defined by a word2vec model trained on the medical records), and return documents with any of the similar terms. Rank documents by the number and extent of similar terms in the document.
- **Learning-to-rank keyword search**: A learning-to-rank system takes a few labeled examples, and then adjusts the similarity weights to prioritize certain terms over others. Documents are ranked like the expanded keyword search system, but with updated similarity weights. Clinicians with different specialties read notes with varying intentions. Learning-to-rank assists researchers identify content that is relevant to their specific problem.

![Figure 5: An example of expanded search terms for ‘diabetes’.

In addition to search, we have found that highlighting relevant text within notes can help workers more quickly focus on important information. As Figure 4 shows, similar words to the search term are highlighted in dark green, and moderately similar terms are highlighted in yellow. The similarity value is determined by a word2vec model. Because clinical notes often contain many pages of information, highlighting allows workers to scroll to relevant text quickly.

5 DISCUSSION AND NEXT STEPS

We have completed more than six crowdsourcing projects including projects that:

(a) Analyze how well barriers to diabetes care are documented;
(b) Analyze if a patient had dialysis two weeks prior to surgery;
(c) Analyze if a note was relevant to a patient’s diabetes treatment.

For each project, we conducted workshops, and recruited medical students and nursing students to participate in the crowd (over a dozen have participated). We paid the workers a flat fee to complete each project, which was determined by multiplying an hourly rate times the expected number of hours of work.

For many projects, researchers have asked that workers snip the text used to make their decision. These snippets are then provided to an expert for validation. Even though this process requires an expert to review all answers, we find
it is useful as the workers complete the time consuming task of scanning the entire document, while the expert simply reviews and approves snippets. If an expert’s time is limited and much more costly than workers, then this design can be effective.

6 CONCLUSION

In this paper, we presented a crowdsourcing framework for sensitive data sets, such as medical records. We developed a crowdsourcing platform that protects patient privacy and a set of helper libraries to assist workers complete tasks efficiently. Our aim is to help medical researchers attain high quality labels faster and more cheaply than previously possible. Future extensions of the framework include level-of-expertise weighted answers, quorum-detection, and machine learning prediction label assistance.

Acknowledgments: Data was obtained from Vanderbilt University Medical Center’s Synthetic Derivative, which is supported by institutional funding and by the Vanderbilt CTSA grant ULTR000445 from NCATS/NIH.

Funding: Crowd Sourcing Labels from Electronic Medical Records to Enable Biomedical Research Award Number: 1 UH2 CA203708-01

References


Adapting Word Embeddings from Multiple Domains to Symptom Recognition from Psychiatric Notes
Yaoyun Zhang, Ph.D., Hee-Jin Li, Ph.D., Jingqi Wang, M.S., Trevor Cohen, MBChB, Ph.D., Kirk Roberts, Ph.D., Hua Xu, Ph.D.
School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract
Mental health is increasingly recognized as an important topic in healthcare. Information concerning psychiatric symptoms is critical for the timely diagnosis of mental disorders, as well as for the personalization of interventions. However, the diversity and sparsity of psychiatric symptoms make it challenging for conventional natural language processing techniques to automatically extract such information from clinical text. To address this problem, this study takes the initiative to use and adapt word embeddings from four source domains – intensive care, biomedical literature, Wikipedia and Psychiatric Forum – to recognize symptoms in the target domain of psychiatry. We investigated four different approaches including 1) only using word embeddings of the source domain, 2) directly combining data of the source and target to generate word embeddings, 3) assigning different weights to word embeddings, and 4) retraining the word embedding model of the source domain using a corpus of the target domain. To the best of our knowledge, this is the first work of adapting multiple word embeddings of external domains to improve psychiatric symptom recognition in clinical text. Experimental results showed that the last two approaches outperformed the baseline methods, indicating the effectiveness of our new strategies to leverage embeddings from other domains.

Introduction
Mental health is increasingly recognized as an important topic in healthcare. In recent years, there has been rapid growth in the implementation of electronic health records (EHRs), leading to an unprecedented expansion in the availability of dense longitudinal datasets for clinical and translational research for psychiatric disorders. Psychiatric symptoms, as one type of fundamentally important information, are usually obtained through interpersonal communications and recorded in clinical text in EHRs. Information about the nature, severity, and impact of psychiatric symptoms is indispensable for both the diagnosis of mental disorders, and the customization of interventions to treat them. Therefore, it is desirable to develop automated approaches to extract psychiatric symptoms from clinical text.

However, psychiatric symptoms often consist of subjective and individualized descriptions, which are present in details of the patient’s experience (Figure 1). Instead of a single word or simple noun phrase, psychiatric symptoms have tremendous syntactic and semantic variability. Moreover, symptoms in clinical notes of different institutions, different types of mental disorders (e.g., bipolar disorder vs. substance abuse) and different populations (e.g., adults, teenagers, military veterans) may have their own sub-languages. Therefore, it is quite challenging for traditional natural language processing (NLP) techniques to automatically extract such diverse mentions of psychiatric symptoms from text. On one hand, existing clinical terminologies such as in Unified Medical Language System (UMLS), SNOMED-CT, and ICD-9 code have low coverage of such complex expressions. On the other hand, conventional supervised learning-based methods (typically the algorithm of conditional random fields-CRF) are widely employed for clinical concept recognition. Such methods rely heavily on NLP feature engineering, leading to systems not generalizable to other clinical texts with different sub-languages.

Although emerging research activities have used NLP techniques to unlock information in psychiatric text for various applications recently, only a few efforts have been made to extract psychiatric symptoms. Given that the diversity and sparsity of psychiatric symptoms require a vastly larger labeled corpus compared to concepts like diseases and medications, Gorrell et al. applied active learning to alleviate this problem for negative symptom recognition of schizophrenia. However, domain experts still need to be heavily involved in this process. In contrast, our previous work employed an unsupervised framework to address this problem by leveraging distributional representation of phrases. In their work, symptoms were collected from online knowledge sources and candidate symptoms were identified based on semantic similarity. However, our previous study mainly focused on extracting a high-quality candidate list of symptoms; further accurate classification of each specific mention in the context of clinical notes is left to future work. In previous research, we have used distributed representations of phrases composed from word embeddings produced using Latent Semantic Analysis (LSA) as a basis for supervised learning of the relationship between phrases in psychiatric narrative and diagnostically meaningful categories, such as “mood disor-
...“dangerousness”\[12\]. While this work demonstrated the utility of distributed representations learned from a background corpus for the classification of psychiatric symptoms, it was limited in scope to a small number of symptom categories, and did not evaluate the influence of different source corpora on classifier performance.

Neural networks, or deep learning-based methods, are growing in popularity as approaches to NLP. Deep learning-based methods do not need time-consuming and labor-intensive feature engineering\[13,14\]. Instead, word embeddings pre-trained from large-scale unlabeled corpora are usually used as features.\[15\] As the currently most widely-used distributional semantic representation (i.e., vector representation) of words, neural word embeddings (such as those produced by the word2vec software package\[16\]) are assumed to capture the latent syntactic/semantic information of a word, because the resulting vector representations for words will be similar if these words occur in similar local contexts.\[16\] While this is also the case for prior distributional models such as LSA, neural embeddings have shown advantages over other distributional models with optimized parameters in some experiments.\[17\] Deep learning methods are well-suited to leverage these representations, as they both exemplify the parallel distributed representation paradigm.\[18\] Thus, the framework of deep learning-based methods with word embedding features has stronger generalizability to resolve the diversity and sparseness of natural language.\[16\] This is particularly important in psychiatry on account of different ways in which patients describe their experience of illness, which may explain in part the promising performance of deep learning based methods on NLP tasks in this domain. Furthermore, a recent study of Habibi et al. demonstrated that using deep learning-based methods outperformed state-of-the-art entity-specific NER tools and an entity-agnostic CRF implementation by a large margin, by conducting experiments on 33 data sets covering five different entity classes in the biomedical literature and patent domains.\[15\]

However, due to the fact that there is no publicly available, large corpus of psychiatric notes, external resources are necessary for building the distributional representations. On the other hand, word embeddings from multiple large-scale external resources such as MEDLINE and Wikipedia have been applied and demonstrated their effectiveness on clinical NLP tasks, such as estimation of the semantic similarity and relatedness between clinical concepts\[19\], NER\[20\], and assertion identification.\[20\] So far, current works mainly investigated word embeddings derived from external resources by using them directly, or combining the corpora of multiple sources to train a single word embedding model.\[10,19,20\] An issue with the former approach is that word embeddings from different sources may need to be adapted to the clinical domain for optimal performance, and an issue with the latter is that the distributional information from larger general-domain corpora may overwhelm that of a small target domain corpus. Domain adaptation technologies could be a solution to address the first of these problems. Domain adaptation attempts to maximize the use of existing data (source) for the data of interest (target) by learning useful aspects of source data. It has been applied for various NLP tasks in the biomedical domain, such as semantic role labeling in biomedical literature\[21\] and clinical notes,\[22\] automatic discourse connective detection in biomedical text,\[23\] and de-identification of psychiatric text.\[24\] However, few studies have been conducted on domain adaptation of word embeddings, yet.

![Figure 1](image-url)

Figure 1. An example paragraph from psychiatric notes with symptoms. The psychiatric symptoms are highlighted in italic.

This study addresses the use and adaptation of word embeddings derived from external domains for symptom recognition from psychiatrist text, using deep learning-based algorithms. Specifically, four domains – intensive care, biomedical literature, Wikipedia and Psychiatric Forum are used as source domains (Ds) and adapted to the target domain (D\(T\)) of psychiatrist text. Word embeddings of each Ds are investigated using four different approaches. First, two basic approaches commonly used in previous works in the clinical domain \cite{cite} are employed: (1) using word embeddings of Ds only and (2) directly combining data of Ds and D\(T\) to generate word embeddings. In addition, two novel approaches are also implemented to adapt word embeddings of Ds to D\(T\): by (3) assigning different weights to word embeddings of Ds and D\(T\) and (4) retraining the word embedding model of each Ds using the available dataset of D\(T\). To the best of our knowledge, this is the first work of adapting multiple word embeddings of external domains to psychiatric symptom recognition from clinical text. Experimental results demonstrated that the proposed strategies achieve promising performance, which outperformed the baseline of only using word embeddings of D\(T\) and a CRF-based system with fine-tuned features.
Methods

Study design

Figure 2 shows the study design of domain adaptation based psychiatric symptom recognition. For this task, data from four source domains (Ds) - intensive care, biomedical literature, Wikipedia and Psychiatric Forum - are employed for domain adaptation. After word embeddings are generated by applying different strategies on data of Ds and Dτ, they are fed into the deep learning-based methods as initial features to train automate systems for psychiatric symptom recognition. The systems are evaluated based on predictions on the test dataset.

Datasets

We used five unlabeled datasets in our study, four as source domain datasets and one as the target dataset, as described below:

Source domain datasets: (1) MIMIC III\textsuperscript{25} is a publicly available collection of clinical notes in intensive care unit (ICU), which is commonly used to generate word embeddings for clinical NLP. (2) MEDLINE is a collection of scientific article abstracts in the biomedical domain. While MEDLINE covers clinical concepts such as diseases, medication and treatments, it contains well-formed sentences different from the telegraphic style in clinical notes. We used the MEDLINE data of 2013 in this study. (3) Psychiatric Forum is a collection of forum posts on the WebMD Community. The forum posts are written largely by health consumers, who use similar expressions to those of patients, as recorded in their psychiatric notes. (4) Wikipedia is an online encyclopedia built from crowdsourcing. It has extensive coverage of topics of multiple domains, including medical topics. The text in Wikipedia is usually well formed.

Target domain dataset: The psychiatric notes provided by the CEGS N-GRID 2016 challenge\textsuperscript{36} organizers are used for as the target domain dataset in this study. This is the first corpus of mental health records released to the NLP research community, which contains about 1,000 initial psychiatric evaluation records. These records were produced by psychiatrists during the course of the elicitation of psychiatric signs and symptoms, disorders, and other medical conditions in order to decide the course of treatment.

Table 1 displays the statistics of the five corpora, including the size of data, number of unique tokens and the percentage of token overlap between each source and the target data.
Methods of using word embeddings of source domains

Firstly, two basic approaches commonly used in previous works in the clinical domain are employed:

1. **Source_only**: directly using word embeddings of each $D_s$ to initiate the features of the deep learning-based method;

2. **Source+target**: directly combining datasets of each $D_s$ with $D_T$ to generate the word embeddings;

Moreover, two domain adaptation strategies are investigated, in order to better leverage the word embeddings of $D_s$:

3. **Weighted_concatenate**: concatenating the word embedding vectors of $D_s$ and $D_T$. Different weights are assigned to the word embeddings of $D_s$, based on the intuition that word embeddings more important to the system performance should be assigned higher weights as features.

4. **Retrain_source**: re-training the word embedding models of $D_s$ using $D_T$. In this approach, the final input and output weights of a neural network trained on $D_s$ serve as the initial input and output weights for the derivation of embeddings from $D_T$. In this way, the original semantic distributions and representations of words in $D_s$ are adjusted in accordance with their distributions in $D_T$, mediating their adaptation to the target domain.

Deep learning algorithm

The deep learning algorithm of long short-term memory network-conditional random field (LSTM-CRF) proposed by Lample et al. is employed to build the psychiatric symptom recognition model. This algorithm adds a layer of CRF based prediction model on top of the original bi-directional LSTM structure of recurrent neural network (RNN), and has shown state-of-the-art performance in various NER tasks in open domain and biomedical text.

Experiments and evaluation

**Method comparison**: Experiments were conducted to systematically investigate the performance of the Source_only, Source+target, weighted_concatenate, and Retrain_source methods. For the weighted_concatenate method, weight in the range of $[1,10]$ was assigned to word embeddings of $D_s$ and the optimal performance was reported. Moreover, to examine the effectiveness of these algorithms, three baseline methods were also developed for comparison: the “Randomize” method without using any pre-trained word embeddings, the “Target_only” method using only word embeddings of $D_T$ to train a model, and a CRF-based model using a fine-tuned feature set.

**Parameter setup**: The gensim implementation of the neural network architectures provided by the word2vec package [cite] was used to train the word embeddings of each domain, because this package permits retraining an existing model. The corpus used for retraining is not necessarily the same as the corpus from which the original model was generated. The parameters of the word2vec model include: (1) the skip-gram architecture was adopted to train the model; (2) the window size was set to 4; (3) in this preliminary study, the dimensionality of embedding vectors was set to 50 for the Target_only, Source_only, Source+target and Retrain_source methods, and 100 for the Weighted_concatenate method; (4) the initial learning rate (i.e., alpha) was 0.025; (5) and all the words with total

---

**Table 1.** Corpus statistics for MIMIC, MEDLINE, Psychiatric Forum, Wikipedia and psychiatric notes.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size of data</th>
<th>#Unique tokens</th>
<th>#Coverage of target tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>MIMIC III</td>
<td>1.95G</td>
<td>257,472</td>
</tr>
<tr>
<td></td>
<td>MEDLINE</td>
<td>1.3G</td>
<td>251,080</td>
</tr>
<tr>
<td>Psychiatric Forum</td>
<td>78.5M</td>
<td>36,336</td>
<td>3,264 (78.8%)</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>10.4G</td>
<td>2,448,552</td>
<td>4,063 (98.0%)</td>
</tr>
<tr>
<td>Target</td>
<td>Psychiatric Notes</td>
<td>2.4M</td>
<td>4,144</td>
</tr>
</tbody>
</table>
frequency lower than 5 (i.e., min_count) were ignored. Each word embedding model was trained for 100 epochs. As for the LSTM-CRF algorithm, we followed the optimal parameter setup established in the work of Lample et al.\textsuperscript{15}

\textit{Evaluation:} 400 psychiatric notes are annotated with symptoms. These were split into training (60%), development (20%) and test datasets (20%) for experiments. Each model was trained for a total of 100 epochs using the training set. The performance on the test dataset was reported using the optimal model evaluated on the development set. The performance of psychiatric symptom recognition was evaluated using precision, recall and the F-measure.

\section*{Results}

\textbf{Table 2.} Results for RNN-based psychiatric symptom recognition using word embeddings from multiple domains (%).

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Randomize</td>
<td>68.93</td>
<td>63.06</td>
<td>65.87</td>
</tr>
<tr>
<td>Psychiatric Notes</td>
<td>Target_only</td>
<td>70.82</td>
<td>64.05</td>
<td>67.26</td>
</tr>
<tr>
<td>Psychiatric Forum</td>
<td>Source_only</td>
<td>67.85</td>
<td>67.42</td>
<td>67.63</td>
</tr>
<tr>
<td></td>
<td>Source+target</td>
<td>69.34</td>
<td>65.52</td>
<td>67.38</td>
</tr>
<tr>
<td></td>
<td>Weighted_concatenate</td>
<td>72.01</td>
<td>64.05</td>
<td>67.80</td>
</tr>
<tr>
<td></td>
<td>Retrain_source</td>
<td>71.60</td>
<td>64.84</td>
<td>68.05</td>
</tr>
<tr>
<td>Mimic</td>
<td>Source_only</td>
<td>69.66</td>
<td>65.16</td>
<td>67.34</td>
</tr>
<tr>
<td></td>
<td>Source+target</td>
<td>67.86</td>
<td>66.71</td>
<td>67.28</td>
</tr>
<tr>
<td></td>
<td>Weighted_concatenate</td>
<td>69.52</td>
<td>66.90</td>
<td>68.19</td>
</tr>
<tr>
<td></td>
<td>Retrain_source</td>
<td>71.25</td>
<td>67.87</td>
<td>69.52</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>Source_only</td>
<td>69.99</td>
<td>69.04</td>
<td>69.51</td>
</tr>
<tr>
<td></td>
<td>Source+target</td>
<td>69.80</td>
<td>68.82</td>
<td>69.30</td>
</tr>
<tr>
<td></td>
<td>Weighted_concatenate</td>
<td>71.42</td>
<td>66.86</td>
<td>69.07</td>
</tr>
<tr>
<td></td>
<td>Retrain_source</td>
<td>69.50</td>
<td>66.19</td>
<td>67.80</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>Source_only</td>
<td>72.18</td>
<td>63.85</td>
<td>67.76</td>
</tr>
<tr>
<td></td>
<td>Source+target</td>
<td>71.82</td>
<td>64.05</td>
<td>67.71</td>
</tr>
<tr>
<td></td>
<td>Weighted_concatenate</td>
<td>68.95</td>
<td>67.34</td>
<td>68.14</td>
</tr>
<tr>
<td></td>
<td>Retrain_source</td>
<td>73.62</td>
<td>64.81</td>
<td>68.49</td>
</tr>
</tbody>
</table>

Table 2 illustrates the results for RNN-based psychiatric symptom recognition using word embeddings from multiple domains. The Target_only baseline of using word embeddings generated from psychiatric notes outperformed the model with randomized word embedding features (F-measure: 67.26\% vs. 65.87\%). Applying word embeddings of the source domains further improved the performance of the Target_only baseline. For the Source_only method, the Wikipedia dataset yielded the highest F-measure of 69.51\% among the four source datasets, while MIMIC produced the lowest F-measure of 67.34\%.

Interestingly, for this task of psychiatric symptom recognition, the performance of Source+target was lower than the Source-only method. The Source-only method performed best when Wikipedia was the source corpus only. With all other corpora, models that considered the target prevailed. This may reflect the importance of general-domain semantic information for the interpretation of patients’ descriptions of their subjective experience of their symptomatology, a point we have argued previously.\textsuperscript{12,36} Assigning higher weights to word embeddings of the source domains in the concatenation (Weighted_concatenate) generally outperformed the methods of Source+target and Source_only, with the exception of the Wikipedia corpus. Nevertheless, Wikipedia still produced the highest F-measure of 69.07\% when using concatenation methods among the four Ds. In contrast, the Retrain_source method achieved the highest F-measure for three Ds (Psychiatric Forum, MIMIC and MEDLINE), with the optimal performance of 69.52\% yielded by retraining on MIMIC. In contract, Wikipedia produced the lowest F-measure of 67.80\% with this method.
To examine whether our proposed methods could achieve the state-of-the-art performance for psychiatric symptom recognition. We further compared our methods with the CRF algorithm using a fine-tuned feature set. 28 Table 3 listed the performance of using CRF, Target_only, Source_only with Wikipedia as Ds, Weighted_concatenate with Wikipedia as Ds and Retrain_source with MIMIC as Ds. Considering the practical usage of the psychiatric symptom recognition system, performance of both exact match and partial match was reported. As can be seen, the performance of CRF was better than the Target_only method (F-measure: 67.40% vs. 67.26%), demonstrating that CRF was a strong baseline. Notably, by adapting word embeddings from source domains using different methods, the psychiatric symptom recognition systems built in this study achieved better performance than the system based on CRF. Among all the methods, Retrain_Source with MIMIC as Ds achieved the best F-measure of 86.80% in terms of partial match.

Table 3. Performance comparison between the Deep learning based algorithm and the CRF algorithm for psychiatric symptom recognition (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>Exact</td>
<td>75.10</td>
<td>61.14</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>91.90</td>
<td>79.21</td>
</tr>
<tr>
<td>Target_only</td>
<td>Exact</td>
<td>70.82</td>
<td>64.05</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>85.46</td>
<td>84.70</td>
</tr>
<tr>
<td>Source_only (Wikipedia)</td>
<td>Exact</td>
<td>69.99</td>
<td>69.04</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>87.12</td>
<td>86.13</td>
</tr>
<tr>
<td>Weighted_concatenate (Wikipedia)</td>
<td>Exact</td>
<td>71.42</td>
<td>66.86</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>88.72</td>
<td>83.18</td>
</tr>
<tr>
<td>Retrain_Source(MIMIC)</td>
<td>Exact</td>
<td>71.25</td>
<td>67.87</td>
</tr>
<tr>
<td></td>
<td>Partial</td>
<td>86.21</td>
<td>87.35</td>
</tr>
</tbody>
</table>

Table 4. Error analysis of deep learning based psychiatric symptom recognition. False positive/negative errors are highlighted in italic.

<table>
<thead>
<tr>
<th>Error type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive</td>
<td></td>
</tr>
<tr>
<td>Term taken out of context</td>
<td>She felt very well from a mood and anxiety standpoint prior to pregnancy</td>
</tr>
<tr>
<td>Non-specific symptom</td>
<td>It is in context of needing to work long hours and sometimes can be associated with impulsive incidents in the remote past but no recent episodes</td>
</tr>
<tr>
<td>Non-psychiatric symptomatology</td>
<td>Patient is legally blind and has had impaired hearing since birth.</td>
</tr>
<tr>
<td>False negative</td>
<td></td>
</tr>
<tr>
<td>Complex syntactic structure</td>
<td>notices light sensitivity, emotional numbness / detachment / lack of interest in usual activities cluster B traits , ( rigidity , interpersonal difficulty , degree of perfectionism )</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>Denies history of symptoms of AH / VH / TH / OH / GH / TI / TB / TW / IOR</td>
</tr>
<tr>
<td>Rare symptom pattern</td>
<td>At heaviest use was smoking 7 grams cannabis per day ( for 5 mths , fall 2082 while at college )</td>
</tr>
<tr>
<td>Telegraphic writing</td>
<td>Brother bipolar sometimes violent</td>
</tr>
</tbody>
</table>

Discussion

Recently there has been a rapid growth of deep learning methods, and their application to various NLP tasks. This growth has been promoted by the availability of word embeddings pre-trained from large-scale corpora of multiple domains. However, many clinical sub-domains may not have publicly available, large corpus, and external resources
are necessary for building the distributional representations. This study exploits word embeddings of external domains for clinical NLP tasks. Specifically, we take the psychiatric symptom recognition task as a typical use case. Extracting psychiatric symptoms from clinical text suffers from the data sparseness problem and a low coverage of relevant terms in existing biomedical lexicons, making it challenging to apply traditional named entity recognition methods to this task. We evaluated different approaches toward adapting word embeddings from four source domains to recognize entities of psychiatric symptoms using deep learning methods. To the best of our knowledge, this is the first work of adapting multiple word embeddings to psychiatric symptom recognition. Experimental results showed that the proposed domain adaptation strategies could achieve promising performance, outperforming the other strategies evaluated, and a system built with CRF algorithm and fine-tuned features.

**Table 5.** Examples of partial matched psychiatric symptoms.

<table>
<thead>
<tr>
<th>Symptom annotation</th>
<th>Partially matched symptom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicidal thoughts</td>
<td>Periods of suicidal thoughts</td>
</tr>
<tr>
<td>etoh or drug use</td>
<td>drug use</td>
</tr>
<tr>
<td>Complicated Grief; pain medication abuse</td>
<td>Complicated Grief and pain medication abuse</td>
</tr>
<tr>
<td>subjective sense of &quot;brain fogginess&quot;</td>
<td>sense of &quot;brain fogginess&quot;</td>
</tr>
<tr>
<td>had anxiety all my life</td>
<td>anxiety</td>
</tr>
</tbody>
</table>

To further improve the performance, we looked into the current prediction errors of psychiatric symptom recognition. The major reasons of false positive and false negative errors are illustrated in Table 4. A majority of false positive errors are caused by recognizing terms that may stand for psychiatric symptoms in wrong context, such as the term of “anxiety” in the context of “from a mood and anxiety standpoint”. In addition, some general expressions that do not convey specific psychiatric symptoms are another cause of false positive errors. Moreover, although some authentic clinical symptoms/disorders are recognized, there is no explicit evidence in the context to indicate that they are related to mental disorders, which leads to false positive errors. As for the false negative errors, psychiatric symptoms present in complex syntactic structures such as conjunctive structures are often missed to be identified. Besides, abbreviations of psychiatric symptoms are frequently used in psychiatric text and miss-identified by the system. Some psychiatric symptoms are of rare patterns, whereas the telegraphic writing style in psychiatric text also causes some false negative errors. Furthermore, we also reviewed the symptom predictions partially matched with gold-standard symptoms; some typical examples are illustrated in Table 5. As can be seen, some of the boundary errors were not critical, in terms of recognizing essential symptoms entities. Therefore, we argue that in-exact matching could be reasonable in psychiatric symptom recognition.

Previous works on exploiting word embeddings of multiple Ds have obtained different findings in terms of their contribution to specific tasks in $D_T$. For example, Roberts$^{26}$ assessed the performance of word embedding based features on the i2b2 2010 concept recognition and assertion identification tasks and found that merging multiple corpora to generate word embeddings generally worked best. He also pointed out that the single-best corpus was generally task-dependent. Pakhomov et al.$^{25}$ constructed word embeddings of clinical terms for semantic similarity and relatedness between clinical concept pairs and found that measures computed from biomedical literature were on par with measures computed from clinical reports, while measures from Wikipedia were worse than sources from the biomedical domain. Zhang et al.$^{31}$ used embeddings of short text and semantic similarity to identify psychiatric symptom candidates and found that the dataset from Psychiatric Forum contributed the most and the MIMIC corpus decreased the performance when merged with the other corpora. In contrast, our study of using word embedding based features and deep learning-based supervised methods for psychiatric symptom extraction found that the two novel methods of Weighted_concatenate and Retrain_source achieved better performance in general, validating the effectiveness of our proposed approaches. Moreover, directly merging corpora of Ds and $D_T$ (i.e., Source+target) decreased the performance slightly. Notably, the performance when using Wikipedia was an exception from MIMIC, MEDLINE and Psychiatric Forum in that the Source_only method performed the best, while Retrain_source got the lowest F-measure of 67.78%, a marked contrast to its leading performance with all other corpora.

Limitation and future work: One limitation of our current work is that the performance difference of varying dimensions of word embedding vectors has not been examined. In addition, psychiatric symptom lexicons and syntactic patterns commonly present in the target domain will be incorporated as features of the deep learning methods, in order to tailor the system to the target domain. Besides, as shown in Table 3, the CRF-based system got higher precision than deep learning based systems, while deep learning based systems usually obtained relatively higher recall.
Therefore, we will explore potential approaches to leverage the advantage of these two types of algorithms in a single system. Furthermore, domain adaptation strategies to adapt word embeddings from other domains as features in CRF will also be investigated.

**Conclusion**

This study takes the initiative to use and adapt word embeddings of four source domains – intensive care, biomedical literature, Wikipedia and Psychiatric Forum – to recognize symptoms in the target psychiatric domain using deep learning-based methods. Experimental results showed that the proposed domain adaptation strategies achieve promising performance, indicating that deep learning-based methods leveraging word embeddings of source domains have the potential to resolve the data diversity and scarcity challenged of clinical NLP tasks.

**Acknowledgement:** This study was supported in part by R00LM012104-02.

**Reference**

4. Lafferty J, McCallum A, Pereira FC. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML '01; 2001; San Francisco, CA, USA.
24. Hee-Jin L, Yaoyun Z, Kirk R, Hua X. Leveraging existing corpora for de-identification of psychiatric notes using domain adaptation AMIA; 2017; DC.
Identifying Diagnostic Paths for Undifferentiated Abdominal Pain from Electronic Health Record Data

Yiye Zhang, PhD, MS\(^1\), Rema Padman, PhD\(^2\), Paul Epner, MBA, MEd\(^3\), Victoria Bauer, BS\(^4\), Anthony Solomonides, MSc\(^4\), Goutham Rao, MD\(^5\)

\(^1\)Weill Cornell Medical College, New York, NY; \(^2\)Carnegie Mellon University, Pittsburgh, PA; \(^3\)Society to Improve Diagnosis in Medicine; \(^4\)NorthShore University Health System, Evanston, IL; \(^5\)Case Western Reserve University School of Medicine, Cleveland, OH

Abstract

The diagnostic process is a complex, uncertain, and highly variable process which is under-studied and lacks evidence from randomized clinical trials. This study used a novel visual analytics method to identify and visualize diagnostic paths for undifferentiated abdominal pain, by leveraging electronic health record (EHR) data of 501 patients in the ambulatory setting of a single institution. A total of 63 patients reached diagnoses in the study sample. We illustrate steps in identifying diagnostic paths of the study patients, both individually and collectively, and visually present the diversity in their diagnostic processes. Patients in whom diagnoses were obtained generally had more clinical encounters and health services utilization, although their diagnostic paths were more variable than those of the undiagnosed group. The capability of identifying diagnostic paths demonstrated from this study allows us to study larger data sets to determine diagnostic paths associated with more timely, accurate, and cost-efficient diagnosis processes.

Introduction

Diagnosis is an inherently complex, uncertain, and highly variable process\(^1\). Variation in diagnostic approaches can be attributed to different patient characteristics, different disease presentations, and system-related factors such as the availability of specific tests\(^2\). A crucial part of the variation, however, is associated with differing physician heuristics, experiences, and knowledge\(^2\). Similar to widespread variation in treatment patterns, which is frequently associated with high costs and suboptimal outcomes, variation in diagnostic processes is often associated with delayed or incorrect diagnosis, unnecessary expense, over and under treatment, and genuine harm\(^3\). Diagnostic error is the top reason for malpractice claims in the outpatient setting. Roughly 12 million American adults are affected by diagnostic errors annually in ambulatory settings alone\(^4\). A 2006 study of claims involving missed or delayed diagnosis revealed that 59% of diagnostic errors were associated with serious harm and 30% were associated with death\(^5\). The root causes of diagnostic error are complex and multi-factorial, and many errors are the result of both system-related and cognitive factors\(^5\). In an analysis of 100 cases of diagnostic errors, cognitive factors, such as faulty knowledge, and premature closure contributed to diagnostic error in 74% of cases\(^5\). A more recent study by Zwann et al. of 7,926 patient records revealed that human causes, especially faulty knowledge, contributed to 96.3% of cases where a diagnostic adverse event took place\(^6\). Among the greatest diagnostic challenges are “undifferentiated” complaints, which are non-specific symptoms that have not yet manifested into identifiable illnesses\(^7\). These include undifferentiated abdominal pain, dizziness, fatigue, and fever without an obvious source\(^8\). Undifferentiated complaints are associated with especially significant variation in physician diagnostic practices. A survey of primary care physicians found that undifferentiated complaints were the presenting complaint in 64% of cases in which physicians recalled making a diagnostic error\(^9\).

Guidance for systematic approaches to diagnosis, especially in ambulatory settings, is scarce and often of poor quality\(^5\). By contrast, the majority of treatment recommendations are based on randomized trials and high quality systematic reviews\(^10\). Relatively few clinical practice guidelines address diagnosis specifically\(^11\). Diagnostic recommendations in guidelines that address diagnosis and evaluation alone or diagnosis and treatment, are often based on expert consensus or similar weak levels of evidence\(^11\). For example, the majority of recommendations in a recently updated guideline for fever of uncertain source in infants 60 days of age or less, for example, were based on either weak level of evidence or consensus opinion only\(^9\). Therefore, there is a significant gap in knowledge and evidence for the optimal diagnostic strategies, which have major consequences on early identification of diseases and proper referrals to specialists.
In order to identify diagnostic steps that lead to more timely diagnosis, greater accuracy, and more efficient use of diagnostic resources, we must obtain a clear understanding of the current practices. We previously defined a **diagnostic path** to be “the steps taken for diagnostic evaluation of a complaint from initial presentation until either a diagnosis is achieved or the patient and/or physician choose to end the evaluation without obtaining a diagnosis”\(^7\). A path begins with a clinical evaluation, which includes a detailed history and in most cases, a physical examination. This is followed by any number of possible steps, including laboratory or imaging studies, referrals, observation, follow-up visits, or a trial of therapy. A referral may result in additional steps being added to the path. In this study, we define “reaching a diagnosis” as having a diagnosis recorded in the medical record on a minimum of 2 encounters for the complaint in question, and which becomes the basis for further management. A diagnosis must also be different from the chief complaint. For instance, “low back pain” may be a chief complaint, and may also be used as an interim diagnosis by the physician. However, the diagnosis must be more specific than “low back pain” and could include, for example, “lumbar disc herniation.” Alternatively, the diagnostic path may end if the patient or physician chooses to stop the evaluation without achieving a stable diagnosis. This may occur because the patient’s symptom resolves prior to reaching a diagnosis. The patient may no longer seek evaluation for the complaint, even though the symptom may continue to persist, or the patient may be lost to follow-up. The physician may recommend not pursuing the evaluation for a number of reasons such as harmful further diagnostic testing and initiate empiric treatment instead. Thus, a path may be brief, or lengthy and complex, depending on individual cases.

Understanding diagnostic paths for an individual patient or for a large number of patients in a way that is meaningful to clinicians presents a significant challenge. Previously, the only method to identify diagnostic paths was either prospectively, by observing and recording steps taken as patients go through the diagnostic process or through analysis of paper-based charts, which is cumbersome and imprecise. With the availability of Electronic health records (EHRs), we now are presented with an extremely useful opportunity to identify and analyze diagnostic paths using data-driven approaches. Yet, EHR data are heterogeneous, complex, and usually not collected for large-scale analysis of decision making. To address this challenge, we applied process mining techniques combined with visual analytics\(^12\), which is the combination of advanced statistical analytics and visualization techniques for evidence discovery from data\(^13,14\). We believe analysis based on visualization of large number of paths can provide insights that are complementary to a purely quantitative approach.

Process mining, originated from business process management\(^15\), has been used in recent years across clinical domains to identify time-series patterns from data\(^16-18\). Visual analytics, when applied to individual diagnostic paths or paths from thousands of patients, also has promoted easier consumption of data and knowledge discovery such as EventFlow\(^10\) and Harvest\(^20\). One common challenge from previous literature on process mining and visual analytics has been the complexity due to events that share the same timestamps\(^16\). In this paper, we apply a previously developed methodology, which overcomes such complexity, to the construction of diagnostic paths from time stamped EHR data. This methodology has been applied to investigate patterns of care for patients with chronic kidney disease\(^21,12,13\) and left-ventricular assist device (LVAD) implantation\(^22\). We aim to show that this methodology, with appropriate technical refinements, can provide a visual representation of the common diagnostic paths of a cohort, as well as an individual diagnostic path of a patient. Both can be highly informative for clinicians and patients in understanding the diagnostic process. Given the scarcity of the current research into best diagnostic practices, this preliminary study can potentially lead to future research that shift current research and clinical practice paradigms, and establish the foundation for a new type of diagnostic guidance for these types of common, challenging, undiagnosed complaints. To the best of our knowledge, this work is the first to aim to identify and learn diagnostic paths from EHR data.

This paper is organized as follows. Details of the visual analytics methodology and its application to diagnostic path are described in Methods. Data section describes the patient sample characteristics. Preliminary insights and illustration of diagnostic paths are presented in Results. Finally, we conclude with study limitations, extensions, and conclusions in the Discussion section.
Method

There are 3 essential steps in visualizing diagnostic path from EHR data through visual analytics: transformation of the data, construction of common paths, and visualization. With these steps, we will instantiate the abstract flow diagram of the diagnostic process by defining key activities during encounters to generate event sequences representing each patient’s actual encounters.

Data Transformation: The purpose of this step is to reduce the complexity of the data documenting different aspects of the diagnostic process in an ambulatory setting, such as having multiple clinical events documented with the same-day time-stamp. The complexity in data is addressed by developing a novel representation for the clinical data structure, shown in Figure 1, that summarizes the co-progression of all information into a one-dimensional path of chronologically sequenced events for each patient. The flexibility of the methodology allows these data categories to be altered and expanded depending on the availability of appropriate data and variables. Data transformation is applied to all study patients’ records, and each patient’s data is summarized into one and only one path. This step prepares the data for the application of a variety of existing sequential data analytic methods, which is a crucial part of visual analytics.

For instance, if patient 1 has a sequence:

\[ \text{clinical evaluation} \rightarrow [\text{clinical evaluation} + \text{PPI}] \rightarrow [\text{clinical evaluation} + \text{referral to GI}] \rightarrow [\text{clinical evaluation} + \text{EGD}] \]

and, patient 2 has a sequence:

\[ [\text{clinical evaluation}] \rightarrow [\text{clinical evaluation} + \text{referral to GI}] \rightarrow [\text{clinical evaluation}] \]

then, their LCS is of length 2, being:

\[ [\text{clinical evaluation}] \rightarrow [\text{clinical evaluation} + \text{referral to GI}] \]

Then their LCS distance (dLCS) is:

\[ dLCS = |\text{sequence 1}| + |\text{sequence 2}| - 2 \times \text{LCS} = 4 + 3 - 2 \times 2 = 3 \]

Subsequently, in order to identify common patterns in the diagnostic process, we first elicit all transitions seen in the diagnostic paths of all patients. This is performed for each subgroup identified earlier. In a diagnostic path, we refer to an encounter occurring first as source and the successor as target. For example, given a path with encounter
Table 1. Characteristics of patients in the diagnosed and undiagnosed groups

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Diagnosis (N=63)</th>
<th>No Diagnosis (N=438)</th>
<th>P-value for significance of difference across 2 groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment duration in days</td>
<td>585.8±239.45</td>
<td>507.3±290.13</td>
<td>0.4769</td>
</tr>
<tr>
<td>Age (mean, SD)</td>
<td>53.5±16.95</td>
<td>51.4±15.88</td>
<td>0.332</td>
</tr>
<tr>
<td>BMI (mean, SD)</td>
<td>27.7±5.51</td>
<td>27.6±5.57</td>
<td>0.897</td>
</tr>
<tr>
<td>Sex – Female (%)</td>
<td>49.2%</td>
<td>50.0%</td>
<td>1.00</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>African American (%)</td>
<td>1.6%</td>
<td>3.2%</td>
<td>0.900</td>
</tr>
<tr>
<td>Asian (%)</td>
<td>3.2%</td>
<td>5.5%</td>
<td></td>
</tr>
<tr>
<td>White Caucasian (%)</td>
<td>66.7%</td>
<td>63.7%</td>
<td></td>
</tr>
<tr>
<td>Other (%)</td>
<td>28.6%</td>
<td>27.6%</td>
<td></td>
</tr>
</tbody>
</table>

Visualizations: Visualizations are prepared by taking the common transitions extracted using methods from above. Diagnostic paths are represented in graphical form using nodes and edges. Nodes represent single encounters, and edges represent continuation of one encounter to the next. We adjust for path characteristics such as encounter type and frequency using colors and sizes of nodes and edges, described in detail as we present the results in the Results section. We used Gephi 0.9.1, an open-source software for graph and network analysis that uses a 3D render engine to display large networks. The diagnostic paths were plotted first using Gephi’s built-in ForceAtlas2 algorithm, which is scaled for small to medium-size networks and suitable for qualitative interpretation. Following the layout algorithm, we manually adjust the path visualization to clarify the start of all diagnostic paths to various routes before diagnoses. Due to the high variability and thus large number of nodes expected from our data, we chose to use network graphs over other visualization format such as Sankey diagram and algorithms such as LifeFlow for its compactness and interpretability.

Data
Hispanic (%) | 9.5% | 8.0% | 0.617  
Patient’s preferred language  
English (%) | 96.8% | 95.9% | 1.00  
Medical History  
Any allergy (%) | 47.6% | 43.8% | 0.784  
Any past medical hist. on problem list (%) | 69.8% | 66.2% | 0.544  
Any past surgical hist. on problem list (%) | 60.3% | 58.0% | 0.891  
Any family medical history (%) | 69.8% | 66.2% | 0.773  

We obtained EHR (Epic Systems, Verona, WI) data from ambulatory settings from January 2010 to December 2012. We used a small database of records from 501 adult patients with abdominal pain from a single institution in Illinois. All patients presented a chief complaint of new onset of abdominal pain, stomach pain, or epigastric pain, and their data were extracted from the available structured chief complaint fields in the EHR. Based on consensus of the research team, only structured field data relevant to abdominal pain was extracted, including the most common patterns of care: referrals, orders including diagnostic tests and procedures, medication prescriptions, follow-up intervals, and International Classification of Diseases (ICD)-9 codes for diagnoses associated with abdominal pain. Table 1 describes the characteristics of patients and treatments in the study data by patients who received diagnoses and those who did not. We applied t-tests to compare the significance of differences across the 2 groups and listed the p-values for each characteristic. No significant differences were observed.

Table 2. Data element included in the diagnostic paths

<table>
<thead>
<tr>
<th>Category (N)</th>
<th>Classes (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Referral (19)</td>
<td>Addiction (1), Allergy (14), Cardiology (31), Endocrinology (18), Gastroenterology (113), Gynecologic oncology (92), Hematology (11), Nephrology (14), Neurology (42), Neurosurgery (9), Nutrition (10), Oncology (17), Otolaryngology (49), Pain (8), Psychiatry/Psychology (21), Pulmonary (17), Surgery (53), Radiation Oncology (17)</td>
</tr>
<tr>
<td>Order (23)</td>
<td>Albumin (3), Allergen-wheat (2), Bilirubin (1), C. Difficile (43), CBC (1134), Celiac (3), Chemistry (1174), Colonoscopy (95), C-reactive protein (16), E. Coli (2), EGD (53), Fecal Blood (7), Hepatitis (140), H. Pylori (69), Imaging (317), Lipase (124), Pathology (10), Nutrition consult (3), Stool (43), Urinalysis (535), Urology consult (3)</td>
</tr>
<tr>
<td>Medication (10)</td>
<td>Analgesics (892), Anti-infective agents (892), Cardiovascular agents, Central Nervous System Drugs (1111), Endocrine &amp; Metabolic Drugs (863), Gastrointestinal Agents (634), Genitourinary Products (610), Nutritional Products (165), Respiratory Agents (522), Others (1449)</td>
</tr>
<tr>
<td>Diagnosis (noted at least twice)</td>
<td>Anal fissure and fistula (2), Calculus of kidney and ureter (5), Cholelithiasis (4), Diverticulitis of colon (4), Dyspepsia and other specified disorders of function of stomach (3), Esophageal reflux (5), Intestinal infection due to clostridium difficile (2), Unspecified gastritis and gastroduodenitis (2), Other disorders of urethra and urinary tract (14)</td>
</tr>
</tbody>
</table>

Due to the limited size of the study data, data elements were summarized into related classes by the research team. For example, both ‘US Transvaginal Screen and ‘Ultrasound Abdomen or Pelvis’ are summarized as imaging orders. We also performed chart reviews to validate the relevance to abdominal pain of specific diagnostic steps. For example, abdominal ultrasound was always relevant to abdominal pain. Table 2 describes data elements in diagnostic paths that were analyzed.

**Results**

Characteristics of the diagnostic paths in the diagnosed and undiagnosed groups, including encounters before and after reaching diagnoses are shown in Table 3. Using elements listed in Table 2, we found 490 distinct diagnostic paths out of a total of 501. There was a total of 1107 distinct encounters, consisting of different combinations of referrals, orders, medication prescriptions, and diagnoses (or no diagnoses). As the table shows, despite the small size of the data (501 patients), there is tremendous diversity in the diagnostic paths. For example, the average LCS distance is 60.6 and 37.4 in the 2 patient groups, respectively. Larger LCS distance suggests more variable
diagnostic paths, both in terms of encounter content and path length. Therefore, the diagnosed group is more variable compared to the undiagnosed group, and it is also evident from the ranges of path length. The diagnosed group’s path ranges from 4 to 215 encounters, whereas the undiagnosed group’s path ranges from 2 to 146.

Table 3 also lists most frequent orders, referrals, and medications taken by patients in the 2 groups, and the average number of observation per patient. Patients who received diagnoses have had consistently more services across orders, referrals, and medications. The most common 5 orders are the same in both diagnosed and undiagnosed groups, while in different order, and the diagnosed group received larger number of orders compared to the undiagnosed group. For referrals, both Gynecologic oncology and Gastroenterology are among 2 of the most common. Due to the smaller size of the referrals, we only listed the 3 most common referrals. The diagnosed group has surgical referral, and undiagnosed group has Otolaryngology referral as the other. Similarly, for medications, diagnosed group has Genitourinary Products, whereas the undiagnosed group has Endocrine & Metabolic Drugs, as one of the most common 5 medications, respectively.

**Table 3.** Characteristics of the diagnostic paths in the diagnosed and undiagnosed groups, including encounters before and after reaching diagnoses

<table>
<thead>
<tr>
<th></th>
<th>Diagnosed (N=63)</th>
<th>Undiagnosed (N=438)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of distinct paths</td>
<td>63</td>
<td>427</td>
</tr>
<tr>
<td>Min, mean, max number of encounters in paths*</td>
<td>4, 43.2, 215</td>
<td>2, 26, 146</td>
</tr>
<tr>
<td>Number of unique encounter types</td>
<td>505</td>
<td>936</td>
</tr>
<tr>
<td>Number of unique transitions from one encounter to the next</td>
<td>1647</td>
<td>5215</td>
</tr>
<tr>
<td>Average LCS distance (SD)</td>
<td>57.5 ± 45.03</td>
<td>37.4 ± 28.56</td>
</tr>
<tr>
<td>Most frequent orders (average number of observations per patient)</td>
<td>CBC (3.9), Chemistry (3.7), Urinalysis (2.3), Imaging (1.4), Hepatitis (0.5)</td>
<td>Chemistry (2.2), CBC (2.0), Urinalysis (0.9), Imaging (0.5), Hepatitis (0.3),</td>
</tr>
<tr>
<td>Most frequent referrals (average number of observations per patient)</td>
<td>Gynecologic oncology (0.5), Surgery (0.3), Gastroenterology (0.2)</td>
<td>Gastroenterology (0.2), Gynecologic oncology (0.1), Otolaryngology (0.1)</td>
</tr>
<tr>
<td>Most frequent medications (average number of observations per patient)</td>
<td>Other (4.8), Central Nervous System Drugs (2.9), Genitourinary Products (2.7), Anti-Infective Agents (2.6), Analgesics &amp; Anesthetics (2.6)</td>
<td>Other (2.6), Central Nervous System Drugs (2.1), Endocrine &amp; Metabolic Drugs (1.7), Anti-Infective Agents (1.7), Analgesics &amp; Anesthetics (1.7)</td>
</tr>
</tbody>
</table>

Figure 2 displays a visualization of the diagnostic paths of all 63 patients containing all transitions of encounters until reaching stable diagnoses, but not after. In the figure, each node, except for the green node, represents an encounter, and labels on the node show actions taken by providers during the encounter, such as placing a medical order or making a referral. We color-coded the node by index visit, encounters after the index visit, and diagnoses. Since there can be multiple events taking place during one encounter, as reflected in the over 500 unique encounter types, we did not color-code nodes beyond the 3 types. The green node signals the start of each diagnostic path and is not an encounter; purple nodes are an encounter without diagnoses; and the orange nodes are encounters where at least one diagnosis was reached. Size of the nodes and labels represent commonality of the encounter represented by the nodes and labels. Edges thickness represents the number of patients who have experienced the transitions; the thicker the edge, the more patients who have experienced the transition. Also, purple edges represent transitions of encounters before reaching diagnoses, and orange edges represent transitions leading to at least one diagnosis. Edge lengths are determined for layout purposes only and are not representative of any characteristic of the paths. While both Table 3 and Figure 2 show the diversity that exists in the data, the diagnostic paths visualized in Figure 2 more clearly show the dramatically diverse processes that patients go through via the differing number of evaluations, from 1 up to 127 encounters, before reaching stable diagnoses.
Figure 2. Diagnostic paths of all 63 patients with abdominal pain who reached diagnoses.

Figure 3. Eliciting only common (experienced by at least 2 patients) diagnostic paths of 63 patients with abdominal pain who reached diagnoses.

On the other hand, Figure 3 shows the common paths encountered by the 63 patients who received at least one diagnosis. Compared to Figure 2, Figure 3 filters out edges that have edge frequency that is fewer than 2, thereby reducing the complexity seen in Figure 2 and allows easier interpretation of the diagnostic paths. Due to the small sample size, we only used frequency as the threshold and not weight. It is much easier to see in Figure 3 that some patients received their diagnosis after completing imaging orders, whereas others received their diagnosis after 11 encounters with only evaluations and no other actions. However, we did not find significant associations among
specific diagnoses and time of encounter, or previous clinical events. Ten patients reached diagnoses at their initial encounters, including Other symptoms involving abdomen and pelvis (3 patients), Esophageal reflux (2 patients), Diverticulitis of colon (2 patients), Other and unspecified noninfectious gastroenteritis and colitis (1 patient), Disorders of menstruation and other abnormal bleeding from female genital tract (1 patient), Infectious diarrhea (1 patient), Dyspepsia and other specified disorders of function of stomach (1 patient), Calculus of gallbladder without mention of cholecystitis (1 patient), Inguinal hernia, without mention of obstruction or gangrene, unilateral or unspecified (1 patient). Diagnoses obtained after 11th encounter with evaluation only include Urinary tract infection (1 patient), Regional enteritis of unspecified site (1 patient). Diagnoses after performing imaging include Calculus of gallbladder without mention of cholecystitis (1 patient), Diverticulitis of colon (1 patient), and Other calculus in bladder (1 patient).

Figure 4. Selected simplified diagnostic paths among 56 patients with abdominal pain who reached a diagnosis. Lengths of arrows represent time duration between 2 encounters. Dashed arrows indicate omitted encounters; Gray=clinical encounter only with no orders; Green=referral; Blue=diagnostic order; Orange=medication; Purple=order+medication; Yellow=referral+order+prescription; Orange square = stable diagnosis obtained

Furthermore, to illustrate temporality, Figure 4 displays simplified diagnostic paths among a subset (56 patients out of 63 patients) who reached diagnoses. The paths are constructed by summarizing events in Table 2 by their categories. We excluded 7 patients from this figure whose paths were distinctly different from others, and whose number of encounters ranged from 9 to 87, which were too long and complex to fit appropriately in the figure. Unlike Figures 2 and 3, the edges in Figure 4 represent relative duration in time between 2 encounters. Time durations are calculated as average time in days when there is more than one patient. Hence, we see that while 10 patients received diagnoses at their first encounter, there is one patient who received a diagnosis 137 days after having medication prescriptions, and another 5 patients who received diagnoses 18 days, on average, after receiving orders and medication prescriptions. The dashed edges refer to omitted encounters to minimize complexity. For example, the diagnostic path at the very top are experienced by 24 patients who received diagnoses after different number of encounters with evaluations only.

Discussion
In this preliminary study, we aimed to demonstrate that diagnostic paths can be learned and visualized from EHR data. We found tremendous amount of diversity in the diagnostic paths for patients who all presented undifferentiated abdominal pain in a single institution’s ambulatory setting. In this paper, due to the small sample size, our goal was not to make any conclusions about practices that lead to, or not lead to diagnoses. Yet, we did find
that patients who received diagnoses received more clinical services including laboratory orders, referrals, and medication prescriptions. We expect that future studies with a much larger sample size will lead to discovery of diagnostic paths which allow us to conclude associations of evaluation processes with greater timeliness and accuracy in reaching diagnoses, as well as efficient use of clinical resources.

One limitation of the study is the completeness of the data. For example, 3 patients received diagnoses after receiving an imaging order. However, patients in the undiagnosed group received imaging orders for 231 times without reaching diagnoses, as indicated in Table 3. Our data for this study does not include sufficient information about the results of the imaging orders for us to understand whether some patients failed to receive diagnoses because of the inconclusiveness of the imaging results, or whether diagnoses occurred after our study period. Similarly, we do not have results for the laboratory orders listed in Table 2. Therefore, while it is very likely that diagnoses are made based on laboratory results, we are not able to identify such relationships due to the lack of data on laboratory results. If patients were to visit providers from outside institutions, we also would be missing those encounters. These data limitations may explain why we observed a very wide range of time it took before reaching diagnoses. For example, we found that a patient had 127 encounters with evaluation only and no other actions, before reaching diagnoses. In the undiagnosed group, a patient has had 77 encounters with evaluations only, 18 orders of CBC and Chemistry, and 3 imaging orders, without being able to reach any diagnosis. It is valuable to be able to identify patients such as this, whose records likely need detailed review to better understand the reason behind it. When sufficient data are available, rigorously matching patients who received diagnoses versus those who did not, may allow us to discover best practices, as well as exceptions from them, in the diagnostic processes. Moreover, information on providers and patients, such as years of experience, patient demographics, and comorbidities may allow us to classify diagnostic paths by provider and patient types. Furthermore, with longer time period we will identify diagnostic paths which go beyond the first diagnosis to detect changes in diagnoses as more results become available. Advanced analytical techniques for diagnostic paths such as sequential pattern mining can also be applied to generate interesting patterns across subgroups.

Nevertheless, we observed tremendous variability in their diagnostic paths. As shown in Figure 3, filtering out minor events and being able to see the common paths in the population are useful in grasping dominant patterns, but we realize that for specific research questions, low-frequency events need to be evaluated equally to avoid missing important information. Therefore, future work includes developing a visual analytics platform that allows interested users to input EHR from their own practice, and generate diagnostic paths for evaluation, under user-defined thresholds for filtering. More importantly, it is our future aims to understand and reduce the source of variation to only patient-specific factors.

Conclusion
This paper describes a preliminary study using electronic health records (EHRs) data where we, applying a novel visual analytics methodology, examined precisely how a limited sample of patients presenting a chief complaint of new onset of abdominal pain, stomach pain, or epigastric pain, were evaluated before a clear diagnosis was reached/not reached. We aim to demonstrate the potential of innovations in visual analytics for improving the challenging task of diagnostic evaluation for undifferentiated complaints through informatics and advanced visual analytics methods. Leveraging EHR data of 501 patients in a single institution’s ambulatory setting, we examined the diagnostic paths of a total of 63 patients reached diagnoses after a range of number of evaluations, as well as 438 patients who did not reach any diagnosis. We illustrate steps in learning and visualizing diagnostic paths, through which we present the diversity in the diagnostic processes in a more visual way compared to traditional analytical methods. We envision that future work with larger data may lead to identification of common diagnostic evaluation strategies and their associations with timely and accurate diagnoses, as well as efficient use of clinical resources.

References
12. Cook KA, Thomas JJ. Illuminating the path: The research and development agenda for visual analytics. 2005.
Using a Bedside Interactive Technology to Solicit and Record Pediatric Pain Reassessments: Parent and Nursing Perspectives on a Novel Workflow

Raniah N. Aldekhyyel, MHI, PhD(c)¹,4 ; Genevieve B. Melton, MD, PhD¹,² ; Gretchen Hultman, MPH² ; Michael B. Pitt, MD³

¹Institute for Health Informatics, ²Department of Surgery, ³Department of Pediatrics, University of Minnesota, Minneapolis, MN; ⁴King Saud University, Riyadh, Saudi Arabia

Abstract

To measure the impact of a novel interactive inpatient pediatric pain management solution integrating our hospital’s electronic health record system, the nurse communication phones, and the pharmacy dispensing system, we assessed parent and nurse perspectives on the tool’s potential value, benefits, and challenges. A mixed-methods approach with survey instruments containing closed-ended and open-ended questions was administered to 30 parents and 59 nurses (66% and 23% response rate respectively). Overall, parents were more satisfied with the interactive technology experience (90%) compared to nurses (50%) with both indicating timely reassessments of pain being the most valuable feature. Qualitative analysis of nurses’ responses yielded 6 themes for technology benefits and 12 for challenges. While patient-interactive technology solutions appear well-received particularly by parent end-users for pediatric hospital pain management, nurse training and interface improvements may result in higher efficacy, ultimately empowering patients/parents, promoting patient engagement and satisfaction.

Introduction

Healthcare delivery in the United States continues to shift to a patient-centered approach, with increased attention being paid in how to engage patients in the use of their medical data including access to their electronic health record (EHR).¹-⁴ Initially, larger efforts from a policy and EHR vendor perspective have focused on patient-facing health information technology (HIT) tools to provide patients with access to their clinical notes and results from laboratory and imaging tests as well as communicate with their providers and other members of their healthcare team.⁵ Recently, a number of patient-centered HIT tools allow for the capture of patient-generated health data⁶ and outcomes (e.g., home blood pressure or blood glucose levels).⁷-⁹ Ultimately, integrating patient-reported data into the EHR may drive better patient outcomes, assist providers in their clinical documentation practices, and support researchers by improving the quality of EHR data.⁵,⁹ Thus far, however, most of this integration has been limited to outpatient patient/provider systems.¹⁰,¹¹

In the inpatient setting, the use of interactive patient care (IPC) tools accessed at the patient’s bedside including thru the television or other bedside devices is beginning to emerge.¹²,¹³ These IPC tools are often also used as an entertainment platform for patients, by delivering on-demand video and game content, with the added benefit of other virtual patient care interactions such as patient tailored-information delivery through the bedside television (e.g., disease specific educational videos, guided imagery) by the care team selectively assigning content.¹⁴ In most cases, however, IPC tools are stand-alone technology lacking EHR integration limiting the ability to support real-time and meaningful bi-directional interfaces to affect patient care.¹⁵-¹⁶

To improve our ability to provide timely information delivery particularly around pain control, we successfully integrated our inpatient television-based IPC with three other stand-alone HIT systems (the EHR system, the nurse communication phones, and the pharmacy dispensing system) at our children’s hospital, and leveraged this integration to engage parents in their reporting of their child’s pain through a time-triggered television pop-up pain reassessment which was communicated to the nurses’ phone and documented within the EHR.

After implementing this solution, timely documentation of nursing pain reassessment increased 26% compared to the prior year.¹⁷ While this was a statistically significant increase in timely nursing documentation, the overall utilization of the IPC reassessment by patients was low (6.5%). In this study, as this type of patient-centered television-based workflow was previously unreported and the factors contributing to low utilization were unknown, we sought to gain insight from end-users by gathering the perspectives of nurses and parents using this novel interactive pain management tool. This study has the potential to aid in further modifications of the workflow and technology as well as provide broader learnings for improved user experience and utilization for similar IPC tools.
Methods

Hospital Setting and the Pain Management Solution
The University of Minnesota Masonic Children’s Hospital is a 246-bed comprehensive quaternary children’s and mother’s hospital and is part of Fairview Health Systems. In 2014, the hospital piloted a novel pain management solution, which was developed internally to integrate four stand-alone technologies: a television-based IPC system, the EHR system, a nursing call system, and the pharmacy inventory management system. The solution was implemented on four units within the hospital: the intensive care unit, bone marrow transplant unit, and two medical/surgical units that care for a variety of general pediatric and subspecialty patients. The motivations behind developing this system were: (1) to pursue transformative strategies to eliminate pain, (2) activate an additional tool to alert nurses when it was time for pain reassessment and documentation, (3) provide a tool for patients and their families to alert their nurse about the perception of pain and effectiveness of interventions, improve regulatory compliance and efficiency, and (4) reduce variation in access to pain management resources.

Clinical Workflow
Prior to the launch of the new interface, the pain management clinical workflow was mostly manual with no system triggers. The patient/parent reported the level of pain by calling for the nurse, or the nurse identified pain as a problem while interacting with the patient. The nurse was then required to document the pain rating score in the EHR, check pain medication orders, and select and administer the appropriate medication. To conduct patient pain reassessment, the nurse was required to anticipate the medication peak, (which varies for each medication), performs the reassessment, and document the pain rating in the EHR. If pain was still an issue indicated by the patient/parent, the nurse selected an additional intervention (another medication, and/or non-pharmacological support, and/or contact a physician).

After the new interface was implemented, the pain management clinical workflow starts proactively. Once the patient and family are in the admission room, the IPC shows a brief video to set expectations about pain, pain management options, and the partnership between family and the hospital care team in managing pain. When the patient is prescribed pain medications (Figure 1), the nurse first removes the medication from the medication dispensing machine. This initiates a time-based trigger (i.e. 30 minutes for oral pain medications and 15 minutes for intravenous pain medications) for a pop-up window to be displayed on the inpatient television screens showing a pain rating question. Once the patient or parent responds to the question using the television remote or a keyboard, the patient’s pain rating score is communicated to the nurse through their phone and automatically documented in the EHR. Nurses access and interact with the system through their phones and the inpatient computer stations, while patients or parents use the remote bedside clickers or keyboards and their television screens to access and interact with the system.

Figure 1: Overview of Interactive Pain Management Solution Clinical Workflow
Study Design

We used a mixed-method concurrent triangulation approach for this study. Since the experience with the pain management tool differs between nurses and parents, we developed two survey instruments with questions specific to each user’s experience. We built the surveys in Qualtrics@UoM19 and used questions based on examples of other patient and provider technology user experience and satisfaction surveys.20-22 After a pilot study was conducted, surveys were adjusted based on participants’ comments and feedback.

The nurse survey instrument was developed to capture closed ended and open-ended responses. There were 14 closed-ended questions in multiple-choice format designed to collect demographic information, experience with computers, and the perceived usefulness of the tool on a 5-point Likert scale (1=not at all useful to 5=extremely useful). There were 2 open-ended questions designed to capture the experience, benefits and challenges of the tool as perceived by nurses. Nurses were eligible if they had experience with using the solution for at least a month. Invitation emails were sent through the hospital’s nursing email distribution list in December 2015. Reminder e-mails were sent approximately 1, 2, and 3 weeks after the initial e-mail invitation. Upon completion of the survey, nurses were given a $10 gift card.

The patient survey instrument included 32 close-ended questions in multiple-choice format designed to collect demographic information, experience with computers, satisfaction with nursing pain management communication, and the perceived usefulness of the tool on a similar 5-point Likert scale. Based on the feedback received from parents during the pilot study, open-ended questions were omitted from the survey instrument in an effort to minimize time burden of the parents of inpatient children. Parents were eligible if they were 18 years of age or older, used the interactive tool to report their child’s pain for two times or more, and were fluent in English. Parents were surveyed using convenience sampling in which 2 informatics researchers (RA and GH) were notified of inpatient families who met inclusion criteria, and if available and the parents provided informed consent, the researchers administered the survey orally using hand held devices. Participating parents were able to choose from a variety of small gifts valued at under $10 for themselves and their children.

The study was approved by the University of Minnesota Institutional Review Board and the University of Minnesota Masonic Children’s Hospital Nursing Research Council. The study was conducted over a 6-month period starting from December 2015.

Data and Statistical Analysis

The responses to the Likert scale items were analyzed through the Qualtrics website to produce descriptive statistics. Additional analysis was performed in SAS version 9.3 (SAS Institute Inc., Cary N.C.). Non-parametric Spearman correlation, the non-parametric Wilcoxon rank sum test, and the Kruskal-Wallis test were used to assess differences between subgroups. Comparison groups were constructed based on participant demographics with findings being considered statistically significant at P < 0.05. Only correlated items at R > 0.5 are reported in study.

To understand nurses’ responses to the open-ended questions, two reviewers (RA and MP) conducted a thematic content analysis. Reviewers looked for repetition and statements relevant to benefits and challenges with using the solution. Next, the reviewers met together and identified a single set of themes via consensus and created standardized codes for the themes along with a set of definitions. To increase validity and comprehensiveness of the themes, each reviewer independently reviewed the themes while examining the original data. Lastly, each reviewer coded the entire original data independently and then a meeting was convened to reach 100% agreement and consensus between inconsistencies. The final codes were then reviewed and sorted on the basis of the thematic content. The analysis lasted 4 weeks, with 2 group meetings during that period.

Results

Participants

Parents

A total of 30 parents, who met the inclusion criteria were approached to participate during the study period. Twenty parents (66%) agreed to participate. Table 1 summarizes parent demographics.
Table 1. Parent Participant Demographics

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>16 (80)</td>
</tr>
<tr>
<td>Male</td>
<td>4 (20)</td>
</tr>
<tr>
<td>Parent Age (years)</td>
<td></td>
</tr>
<tr>
<td>&lt; 30</td>
<td>5 (25)</td>
</tr>
<tr>
<td>30-39</td>
<td>8 (40)</td>
</tr>
<tr>
<td>40-49</td>
<td>4 (20)</td>
</tr>
<tr>
<td>50-59</td>
<td>3 (15)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Caucasian</td>
<td>14 (70)</td>
</tr>
<tr>
<td>African American</td>
<td>3 (13.0)</td>
</tr>
<tr>
<td>Asian</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Native American</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Education Level</td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>2 (10)</td>
</tr>
<tr>
<td>High school graduate or GED</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Some college</td>
<td>7 (35)</td>
</tr>
<tr>
<td>College graduate</td>
<td>5 (25)</td>
</tr>
<tr>
<td>Postgraduate degree</td>
<td>4 (20)</td>
</tr>
<tr>
<td>Income</td>
<td></td>
</tr>
<tr>
<td>&lt;35K</td>
<td>5 (25)</td>
</tr>
<tr>
<td>35-49.9K</td>
<td>2 (10)</td>
</tr>
<tr>
<td>50-74.9K</td>
<td>4 (20)</td>
</tr>
<tr>
<td>75-99.9K</td>
<td>1 (5)</td>
</tr>
<tr>
<td>100K +</td>
<td>6 (30)</td>
</tr>
<tr>
<td>General level of computer experience</td>
<td></td>
</tr>
<tr>
<td>Less experienced (e.g., browse web, check email, or less)</td>
<td>1 (5)</td>
</tr>
<tr>
<td>Somewhat experienced (e.g., edit photos, use spreadsheet)</td>
<td>12 (60)</td>
</tr>
<tr>
<td>Very experienced (e.g., create web page, write computer programs, or more)</td>
<td>7 (35)</td>
</tr>
<tr>
<td>Patient (child) Age (years)</td>
<td></td>
</tr>
<tr>
<td>&lt;5</td>
<td>5 (25)</td>
</tr>
<tr>
<td>5-11</td>
<td>5 (25)</td>
</tr>
<tr>
<td>&gt;11</td>
<td>10 (50)</td>
</tr>
<tr>
<td>Admission inpatient unit</td>
<td></td>
</tr>
<tr>
<td>ICU</td>
<td>2 (10)</td>
</tr>
<tr>
<td>Medical/Surgical, Hem Onc, Transplant</td>
<td>6 (30)</td>
</tr>
<tr>
<td>Medical/Surgical, Cardiac, Other Specialties</td>
<td>12 (60)</td>
</tr>
<tr>
<td>First admission</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>6 (30)</td>
</tr>
<tr>
<td>No</td>
<td>14 (70)</td>
</tr>
<tr>
<td>Length of Stay</td>
<td></td>
</tr>
<tr>
<td>&gt;3</td>
<td>4 (20)</td>
</tr>
<tr>
<td>3-4</td>
<td>5 (25)</td>
</tr>
<tr>
<td>5-6</td>
<td>3 (15)</td>
</tr>
<tr>
<td>7-8</td>
<td>2 (10)</td>
</tr>
<tr>
<td>&gt; 8</td>
<td>6 (30)</td>
</tr>
<tr>
<td>Prior use of the pain management interactive tool</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>12 (60)</td>
</tr>
<tr>
<td>No</td>
<td>8 (40)</td>
</tr>
<tr>
<td>Number of times the interactive tool was used during the current inpatient stay</td>
<td></td>
</tr>
<tr>
<td>2-3</td>
<td>10 (50)</td>
</tr>
<tr>
<td>4-6</td>
<td>5 (25)</td>
</tr>
<tr>
<td>&gt; 6</td>
<td>5 (25)</td>
</tr>
<tr>
<td>Knowledge of the non-medication patient education resources available through the system</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>12 (60)</td>
</tr>
<tr>
<td>No</td>
<td>8 (40)</td>
</tr>
</tbody>
</table>

*Note: Bold numbers indicate the highest value within a specific group*
Nurses

Although the exact number of email addresses on the hospital’s email distribution list was unknown, there were around 260 registered nurses (RN) and nurse technicians working in the 4 inpatient units included in this study. Fifty-nine nurses (23%) participated in the study. We accepted completed surveys from participants who self-identified as RNs. The survey took an average of 8 min to complete. The majority of nurses indicated they were somewhat experienced with computers (83%, n = 49) such as using spreadsheets, and (80%, n = 47) indicated having more than 6 months experience in using the pain management solution. Nurse participant demographics are provided in Table 2.

**Table 2. Nurse Participant Demographics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>57 (97)</td>
</tr>
<tr>
<td>Male</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>&lt; 26</td>
<td>9 (16)</td>
</tr>
<tr>
<td>26-34</td>
<td>36 (62)</td>
</tr>
<tr>
<td>35-54</td>
<td>11 (19)</td>
</tr>
<tr>
<td>55-64</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Inpatient hospital unit</td>
<td></td>
</tr>
<tr>
<td>ICU</td>
<td>3 (5)</td>
</tr>
<tr>
<td>BMT</td>
<td>13 (22)</td>
</tr>
<tr>
<td>Medical/Surgical, Hem Onc, Transplant</td>
<td>9 (15)</td>
</tr>
<tr>
<td>Medical/Surgical, Cardiac, Other Specialties</td>
<td>34 (58)</td>
</tr>
<tr>
<td>Previous experience with adult patients</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>14 (24)</td>
</tr>
<tr>
<td>No</td>
<td>45 (76)</td>
</tr>
<tr>
<td>General level of computer experience</td>
<td></td>
</tr>
<tr>
<td>Less experienced (e.g., browse web, check email, or less)</td>
<td>8 (14)</td>
</tr>
<tr>
<td>Somewhat experienced (e.g., edit photos, use spreadsheet)</td>
<td>49 (83)</td>
</tr>
<tr>
<td>Very experienced (e.g., create web page, write computer programs, or more)</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Experience working with the pain management interface tool (months)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1 (2)</td>
</tr>
<tr>
<td>3</td>
<td>5 (9)</td>
</tr>
<tr>
<td>4</td>
<td>3 (5)</td>
</tr>
<tr>
<td>5</td>
<td>3 (5)</td>
</tr>
<tr>
<td>&gt;6</td>
<td>47 (80)</td>
</tr>
<tr>
<td>Knowledge of the non-medications patient education resources available through the system</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>37 (62)</td>
</tr>
<tr>
<td>No</td>
<td>22 (37)</td>
</tr>
</tbody>
</table>

*Note: Bold numbers indicate the highest value within a specific group*

**Perceived Usefulness of the Tool**

**Parents**

Most of the parent participants were satisfied with the general experience of using this interactive tool to manage their child’s pain (90%), indicating that it helped their nurse manage their child’s pain in a more timely manner (75%), and many (45%) felt the tool helped them better understand their child’s pain. Additionally, half of the parents indicated the tool led to access to non-pharmacologic alternative resources for pain control including video/visualization resources embedded in the bedside TV entertainment system (Figure 2).

Several differences among parent sub-groups were identified in relation to the perceived usefulness and general satisfaction with the tool. Parent age was found to be negatively correlated with the usefulness of the tool in helping the nurse know the level of child’s pain ($R$=-0.52, $p=0.02$) with younger parents being more satisfied. Results also indicated a correlation between the overall level of satisfaction with the use of the tool and the overall satisfaction with the nursing pain management communication activities. This correlation was found in 5 out of 9 statements that asked parents about levels of satisfaction with specific nursing pain management communication activities: (1) time to discuss concerns in using the tool ($R$=0.61, $p=0.004$), (2) ability of the tool to communicate the level of pain to the nurse ($R$=0.66, $p=0.002$), (3) nurse listening regarding child’s pain ($R$=0.58, $p=0.007$), (4) nurse assessment of
child’s pain (R=0.65, p=0.002), and (5) nurse concern with child’s emotional and physical wellbeing (R=0.60, p=0.005)

**Figure 2:** Perceived usefulness of the tool based on parent responses

**Nurses**
Results from the nursing survey indicated that 50% of the nurses were generally satisfied with the use of the pain management tool, 40% were indifferent, and 10% indicated dissatisfaction. When nurses were asked about the perceived usefulness of the tool, timely reassessment reminders and phone triggers were the top useful features (Figure 3).

We identified several differences among nurse sub-groups with respect to perceived usefulness and general satisfaction with the tool. Nurses with experience working with adults scored lower for reassessment of patient’s pain (p=0.028) and for general satisfaction with the tool (p=0.052). General satisfaction with the tool varied among the hospital units (p=0.034), with the Medical/Surgical, Hematology Oncology, Transplant Unit scoring highest and the Bone Marrow Transplant Unit scoring the lowest.

**Figure 3:** Perceived usefulness of the tool based on nurse responses

**Nurse Perspectives on the Benefits and Challenges of the Solution**
All nurses participating in the study provided a total of 118 responses to the open-ended questions related to the benefits and challenges of using the tool. A total of 45 unique nurse statements were derived and coded from responses to the benefits question, with thematic analysis yielding 2 main themes: (1) nurse benefits and (2) patient benefits (Table 3).
Having phone reminders to reassess patient pain was the most frequently mentioned benefit (n=16) with one nurse stating: “It reminds me to check back with my patients in the appropriate window of time following a pain med (oral or IV)” and another stating, “Sometimes it can be a reminder to check in on patients after giving an oral pain medication; especially if it is within a busy shift.” Several nurses (n=8) cited that the system supported patient empowerment and satisfaction. For example, “…[it] gives the patient/family more control over pain management,” “…there is definitely value in participating in voicing their pain level through this avenue.”

Nurses also described several challenges with using the tool, with 112 unique nurse statements derived and coded from responses to this question. These challenges were mapped to 4 main themes: (1) nurse related, (2) system related, (3) patient related, and (4) organization related.

The most frequent challenge described by nurses (n=16) was uncertainty of patient-rating score. Nurses described concern that patients/parents are using the tool mainly to make the pop-up pain-rating question disappear so that patients continue using the entertainment feature of the system rather than accurately reporting the level of pain. For example one nurse stated, “If patients don’t use it, or just click buttons to get message to go away. Wonder about accuracy in describing pain.” and another stating, “I feel as though most pts just push a button on the GWN to continue watching their program and dont honestly answer at their true pain level”.

The second highest challenge was system related, indicating low utilization of the system among both nurses and patients/parents-unspecified reason (n=15). Some examples of nurse statements include, “…it is very rare that a patient will actually use the Get well network® [IPC vendor name] to reassess their pain.”, and “I don’t see a lot of families clicking the multi-modal pain management strategies button.”

Patient related challenges included low utilization due to patient factors (i.e, age, language, technology comfort) (n=14). Examples of nurse statements include, “Sometimes it is irrelevant because alot of my patients are too young to read.”, and another “It is rarely used properly on the PICU [Pediatric Intensive Care Unit] because our patients are so sick they are unable to utilize the tool.”

**Table 3. Benefits and challenges to the use of the pain management tool**

<table>
<thead>
<tr>
<th>No.</th>
<th>Category</th>
<th>Example</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Benefits</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Nurse benefits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Phone reminder to reassess patient pain</td>
<td>“Super helpful in reminding me to reassess and document...”</td>
<td>16</td>
</tr>
<tr>
<td>1.2</td>
<td>Auto documentation within the EHR</td>
<td>“I like that when the pt responses to their pain after an intervention that it charts it in Epic...”</td>
<td>6</td>
</tr>
<tr>
<td>1.3</td>
<td>Decision support to prioritize patients’ needs</td>
<td>“…If I get a message that pain is increased, I know to prioritize and get back into the room sooner to intervene...”</td>
<td>6</td>
</tr>
<tr>
<td>2.</td>
<td>Patient benefits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.1</td>
<td>Empowerment and satisfaction</td>
<td>“…I think there is definitely value in participating in voicing their pain level through this avenue...”</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>Sense of connection</td>
<td>“…its nice that it lets you know where your patients pain is at after a medication.”</td>
<td>5</td>
</tr>
<tr>
<td>2.3</td>
<td>Non medication resources</td>
<td>“…The best resources in the non-medicaton section of the GWN are videos…”</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td><strong>Challenges</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Nurse related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Uncertain of patient rating scores</td>
<td>“some patients will click the first button they see to get rid of the pop up and it will ding to us that their pain is worse...”</td>
<td>16</td>
</tr>
<tr>
<td>1.2</td>
<td>Less experienced with system /need more training</td>
<td>“Getting all info i need to use it to it's full capability”</td>
<td>11</td>
</tr>
<tr>
<td>1.3</td>
<td>Distraction from other tasks</td>
<td>“… And as a nurse I also find the alerts that come to my phone as annoying. Its just another <em>beep</em> on my phone that distracts me from patient care.</td>
<td>4</td>
</tr>
<tr>
<td>1.4</td>
<td>Discourage best practice - (pain assessment/reassessment documentation)</td>
<td>“…it can never replace the need for a nurse with training a permission to use non-medication pain management....”</td>
<td>2</td>
</tr>
</tbody>
</table>
2. System related

<table>
<thead>
<tr>
<th></th>
<th>2.1 Low utilization (unspecified reason)</th>
<th>“some patients/families do not report the reassessment of pain.”</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.2 Low utilization due to environmental factors</td>
<td>“the biggest challenge is the TV being broken-then I can’t use the interface.”</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>2.3 System design limitations</td>
<td>“Organization of the non-pharmacological pain interventions. Would be easier to have a quick link or to assign these modules…”</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>2.4 Discrepancy pain scales between patient rating and nurse assessment</td>
<td>“…Hence it would be great if the tool used in the GWN could follow the same suit as the one used in Epic.”</td>
<td>3</td>
</tr>
</tbody>
</table>

3. Patient related

|   | 3.1 Low utilization due to patient factors (i.e-age, language, technology comfort) | “some don’t know how to navigate or use it cause they are too young, i see some adults but mostly little kids.” | 14 |
|   | 3.2 Patient annoyed/dissatisfied/uninterested with the tool | “Some patients have expressed annoyance with the question always popping up on their televisions.” | 6 |

4. Organization related

|   | 4.1 Extra work and does not improve overall workflow | “…it doesn’t save us any time or steps.” | 11 |
|   | 4.2 Duplicate charting requirement | “…I wish that there was a way to verify it or co-sign it so that we didn’t have to double chart their response.” | 9 |

The last main theme was coded as being related to the organization and included challenges related to workload and overall clinical workflow (n=11). Nurses described this challenge in statements such as “making sure phones are logged in correctly each shift (one more thing the charge RN has to do)…” and “Honestly, it is one more sign-in to have to do…”.

Discussion

The importance of capturing the perceptions of parents, as consumers, is as important as capturing the perceptions of nurses. Or and Karsh\(^\text{11}\) specifically addressed the main factors that influence consumer acceptance, while other studies examined the effects of introducing health information technology tools on providers.\(^\text{23,24}\) Our objective was to study the perspectives of both parents and nurses towards using an interactive patient care tool in the management of pain at our children’s hospital.

Our main findings indicate cohesive agreement among parents and nurses on the perceived usefulness of the pain management tool. Both indicated that the most usefulness feature of the tool was its ability to support timely reassessment of patients’ pain. Similarly, this feature was the top mentioned benefit that emerged from our qualitative analysis. These findings can be viewed as a factor of successful implementation, as it is aligned with one of the main reasons for implementing this solution. While both parents and nurses indicated that the non-medication pain management resources accessed through the IPC tool was the least useful feature, in comparison to the other system’s features. It is worth noting that this may be a result of a lack of familiarity with this tool. Inexperience with using the full set of features of the solution and the need for more training has emerged as one of the challenges that nurses face with the tool. This is similar to the finding of other research studies examining the challenges to interactive technology adoption.\(^\text{11,25}\)

Although demographic group differences in parents’ perceived usefulness of the tool were not found to be clinically significant, results indicated a correlation between parents’ satisfaction with nursing pain management communication activities and the general satisfaction with the use of the tool. These findings are similar to other findings examining the relationship between that satisfaction with HIT tools and provider communication.\(^5\)

Interestingly, nurses recognized the value of this solution in engaging and empowering patients and their families, increasing patient satisfaction, and creating a communication platform for patients and families to voice their perceptions of pain. This was aligned with the parents’ perspectives, which showed the majority of the surveyed parents satisfied with the experience of using the tool and indicated that it provided a good method to communicate their child’s pain to their nurse.

Many of the challenges that nurses stated suggested a desire for system improvements that would better support nursing pain management care and documentation, rather than resistance to the use of the tool. Nurses indicated uncertainty of the patient pain response scores as the number one challenge. In response to this concern, system
changes were done to the pain rating pop-up question in order to change the default display value “hurts less” to “hurts more”. This change has increased the variability of the patient responses documented in the EHR, which in turn may address this challenge. Educating the patient/parent on the benefits of using the tool and the contribution they make in the pain management process is another vital method that can potentially address this concern.

Low utilization of the tool among patients/parents emerged as a second challenge. This was also found in our previous system’s evaluation study. Similar to other study findings, more education and training is needed to remind nurses about the existing features of the system and how to navigate and educate patients and families about the benefits of using the tool. Other nurses stated challenges related to the duplicate charting requirement and extra work during their shift. Although one of the reasons for developing this solution was to engage patients in their pain management care and support the nursing pain reassessment process, nurses were expecting the tool to replace pain reassessment documentation in the EHR. The overall sense of apprehension among nurses can be partially explained by lack of understanding of the main driver behind implementing this solution. Published literature has highlighted the importance of education for end-users in HIT system implementation, noting that inadequate familiarity and knowledge may potentially lead to user frustration.

Limitations of this study include the relatively small sample size of the participants and our inability to calculate the nurses’ response rate. Although we intended to be as systematic as possible in reaching as many nurses as we could, the exact number of nurse emails were unknown and the perspectives were limited to those nurses who were able to open their work emails on the days and times when messages to participate in the study were sent. Also, the study was limited to one institutional setting; pediatrics. Therefore, future research should explore these findings in an adult setting, where patient factors related to patient age might not be a challenge.

Future studies using time-motion or work sampling techniques may examine the amount of time nurses spend using the solution in their clinical practice and the impact of the tool on time efficiency in documentation practices and patient education. We also plan to conduct a patient/parent usability study to determine what areas can be improved to make the interface easier to use.

Conclusion

To our knowledge, this is the first study to assess parent and nurse perspectives on the implementation of an interactive tool developed to support the pain management clinical workflow within an inpatient pediatric setting. Our results could inform other health care organizations about the feasibility of and potential areas to focus on when implementing an interactive tool integrated with other HIT systems. We found that parents were satisfied with the use of the tool, highlighting its importance as a communication tool with nurses and its effect on timely reassessments of the child’s pain. Nurses recognized the tool’s importance in increasing patient engagement and satisfaction but also expressed some concerns about its validity in reporting patient pain scores. Inpatient interactive tools have the potential to increase patient engagement and communication with clinicians; however, if education and training are not given to end-users, the full benefits of these tools may not be realized. In addition, our findings indicate that nurses are in favor of the solution’s ability to automatically store patient pain reassessments in the EHR, suggesting that future interface changes to the pain rating question may be helpful in supporting pain reassessment documentation practices.

Acknowledgements

This research was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UL1TR000114, the University of Minnesota Clinical and Translational Science Award (8UL1TR000114-02), the University of Minnesota Physicians and Fairview Health Services, and the University of Minnesota Medical Center, Fairview Nursing Research Council. We thank our colleagues Jason Albrecht and Bruce Lindgren who provided insight and expertise that greatly assisted our research.

References

4. Medicare and Medicaid Programs; Electronic Health Record Incentive Program-Stage 3 and Modifications to


23. Meghan Hufstader Gabriel, Emily B. Jones, Leila Samy and Jennifer King Progress And Challenges: Implementation And Use Of Health Information Technology Among Critical-Access Hospitals Health Affairs 33, no.7 (2014):1262-1270


Predicting Mortality in Diabetic ICU Patients Using Machine Learning and Severity Indices

Rajsavi S. Anand1,2, Paul Stey, PhD1,2, Sukrit Jain1,2, Dustin R. Biron1,2, Harikrishna Bhatt, MD1, Kristina Monteiro, PhD1, Edward Feller, MD1, Megan L. Ranney, MD, MPH1,3, Indra Neil Sarkar, PhD, MLIS1,2, Elizabeth S. Chen, PhD1,2
Alpert Medical School1, Center for Biomedical Informatics2, and Emergency Digital Health Innovation Program, Department of Emergency Medicine3, Brown University, Providence, RI, USA

Abstract
Diabetes constitutes a significant health problem that leads to many long term health issues including renal, cardiovascular, and neuropathic complications. Many of these problems can result in increased health care costs, as well risk of ICU stay and mortality. To date, no published study has used predictive modeling to examine the relative influence of diabetes, diabetic health maintenance, and comorbidities on outcomes in ICU patients. Using the MIMIC-III database, machine learning and binomial logistic regression modeling were applied to predict risk of mortality. The final models achieved good fit with AUC values of 0.787 and 0.785 respectively. Additionally, this study demonstrated that robust classification can be done as a combination of five variables (HbA1c, mean glucose during stay, diagnoses upon admission, age, and type of admission) to predict risk as compared with other machine learning models that require nearly 35 variables for similar risk assessment and prediction.

Introduction
Diabetic patients constitute 7% of the United States population, with at least 22.3 million diabetics in the United States today.1 These patients use significantly more healthcare resources than patients with other chronic diseases, accounting for more than 45% of intensive care unit (ICU) patient stays above the age of 65.2 Diabetes itself leads to a higher incidence of nearly all comorbidities including renal, cardiovascular, and neuropathic disease.3 Furthermore, diabetic complications can directly impact how patients will fair in the ICU, with a strong association between diabetes and ICU bloodstream infections established.4 However, the effect of stand-alone diabetes on a patient's risk of ICU mortality has been debated.5 A recent meta-analysis of 141 studies by Siegelaar et al. suggests that having a diagnosis of diabetes in the ICU does not itself directly lead to increased mortality in most ICU settings, only specifically in the cardiac surgical unit.

Management of diabetes in the ICU itself is a controversial topic as tight glucose control in hyperglycemic patients has been seen to have both positive and negative results on mortality outcomes depending on the trial setting.6 The current clinical practice is to have moderate control of blood glucose in the ICU (140 mg/dl), rather than tight control (<110 mg/dl). To date, no study has looked at the combination of glucose control, hemoglobin A1c (HbA1c) values, and comorbidities to predict mortality outcomes. A recent study by Mahmoodpoor et al. demonstrated that higher HbA1c values and admission glucose values were predictive of mortality for ICU patients, showing the value of these variables in predictive modeling for diabetics.7 By better understanding how to treat diabetic patients in the ICU, mortality outcomes could potentially be improved, thus leading to better patient health with less economic burden.

Algorithms exist to predict mortality risk in general patient populations, with the Charlson Comorbidity Index (CCI) and the Elixhauser Comorbidity Index well agreed upon in the literature.8 These two measures calculate risk scores based on the ICD-9-CM diagnosis codes for each patient. The CCI scores are based on the severity of symptoms, weighing more serious conditions with more points; its scoring of general hazard risks have been validated for survival at one and two years.9 The Elixhauser score is based solely on category of comorbidity, and gives one point for each. Furthermore, both of these indices have been applied to ICU patients. However, these indexes have not been validated for diabetics in the ICU. The Diabetes Complications Severity Index (DCSI) has recently been established as an alternative to these measures with a specific focus on diabetic patients using ICD-9-CM diagnosis codes and renal labs.10 The DCSI has yet to be validated in an ICU population of diabetic patients, but it has been shown to be effective in predicting hospitalizations in general diabetic populations. Additionally, this metric has proved successful in predicting which diabetics are most likely to utilize increased healthcare expenditure.11
Machine learning algorithms have been applied to ICU settings, but never specifically to diabetic patient populations in the ICU. Studies have shown the ability to predict risk of mortality in the ICU based on many variables including vitals, labs, surgical history, diagnoses, ventilator settings, and past medical history. Additionally, diabetic specific machine learning has been used in a variety of contexts: detecting hypoglycemia, predicting diagnosis, and further complications, as well as blood glucose classification. \(^{15}\) Recently, there has been success in predicting which diabetic patients are most at risk for specific complications (retinopathy, neuropathy, and nephropathy) with area under the curve statistics of 0.838. \(^{14}\) While useful in predicting complications, neither study investigated diabetic patients in the ICU.

One of the most widely used mortality prediction algorithms in the ICU, the Acute Physiology and Chronic Health Evaluation (APACHE) II algorithm, requires 15 variables (of which 12 are physiological measurements) and does not take into account specific patient populations or diagnoses. \(^{15}\) Subsequent iterations of APACHE (III and IV) have become more and more time consuming, with the need to spend 37.3 minutes on average to enter in relevant information. \(^{16}\) While more accurate at predicting mortality, these algorithms may prove to be unrealistic in their implementation due to the sheer amount of information and time required to perform them. This had led in part to the continued use of the APACHE-II algorithm first described in 1985. Additionally, other scoring algorithms exist for predicting mortality in the ICU such as the Simplified Acute Physiology Score (SAPS) III, which uses 17 variables, including physiological and disease manifestation variables. \(^{17}\) The Sequential Organ Failure Assessment (SOFA) score is also well researched and gives a risk calculation for organ failure. \(^{18}\) Algorithms that require the least possible information upon admission have high practical clinical utility as they allows for rapid implementation and changes in patient treatment strategies. Thus combining diabetic specific metrics and using the fewest possible variables may be of high clinical use in this cohort of patients.

This study aimed to test which of the established comorbidity indices best predicts mortality in diabetic patients in the ICU. By combining other variables of diabetic health including HbA1c, mean glucose during stay, insulin status, as well as these various indices, predictions are made about which diabetics are more likely to have adverse outcomes using both logistic modeling techniques as well as machine learning algorithms.

**Methods**

Data were extracted from a local MySQL database containing Medical Information Mart for Intensive Care III (MIMIC-III) data using SQL queries. Pre-processing (e.g., data recoding) and analyses were done using the Julia programming language (v.0.5). Data were then transformed into various indices and calculations were conducted to create the other variables of interest. The Charlson Comorbidity Index, Elixhauser Comorbidity Index, and the Diabetic Severity Index were calculated as described in the literature, with points assigned for the presence of specific ICD-9-CM codes. Finally, predictive models of mortality were generated on training sets using available Julia packages for statistics and machine learning. Validation occurred on the remaining 30% of available data with results presented from this set of data.

**Data and Variable Selection**

The MIMIC-III database, which includes data from Beth Israel Deaconess Medical Center’s ICU and Hospital from 2001 to 2012, was used for this study. \(^{19}\) This database has comprehensive information regarding ICU admissions and all the data needed to address this question. Data were selected from MIMIC-III detailing admissions, medications, ICU stays, ICD-9-CM diagnoses, and lab values. First, all patients with an ICD-9-CM diagnosis of 250.0 – 259.0 (diabetes inclusive) were selected. A dataset was created that contained the diagnoses, admissions, lab values, and ICU medications for this sample upon which analysis was conducted. Labs of interest were extracted using the following LOINC codes: ‘4548-4’ for HbA1c, ‘2345-7’ for Blood Glucose, and ‘2160-0’ for Serum Creatinine. Generic medication names as well as text encoding ethnicity, insurance, admission type, gender, and admission location were also extracted from the medical record. Patients who did not have HbA1cs or Blood Glucose values listed were excluded.

**Data Processing**

All data processing was conducted using the Julia programming language. The DataFrames.jl package was used to aggregate data in tabular formats. The average blood glucose per stay and comorbidity index scores were calculated for each patient. Demographic data were merged with the lab values and index scores into a single data frame for analysis. Descriptive statistics for the sample were calculated.
Admission type was recoded from the original options (elective, emergent, newborn, or urgent) to give numerical outputs. “Elective” was defined as a previously planned hospital admission, while “emergent” or “urgent” were defined as unplanned medical care. Ethnicity was recoded from 41 categories into numerical outputs. Insurance was also recoded from the original options (government, Medicaid, Medicare, private, or self pay) to give numerical outputs.

Insulin status was determined by cross-referencing patient drug lists with generics that included “insulin” in the name. Blood glucose was averaged for each patient’s stay to create a mean glucose variable. HbA1c was averaged for each patient’s history to gauge overall diabetic health. The former predicts how the patient’s diabetes was handled during the stay while the later was a measure of overall diabetic health over their history.

A DCSI score (0-13) was calculated for each patient as described in the literature using diagnosis codes and serum creatinine levels. Diagnosis codes were categorized into seven different comorbidity categories: (1) ophthalmic, (2) nephropathy, (3) neuropathy, (4) cerebrovascular, (5) cardiovascular, (6) peripheral vascular disease, and (7) metabolic. Depending on the presence or absence of a diagnosis code in a patient’s record, a value was assigned: 0 for no comorbidity, 1 for minor, and 2 for major. Scores were assigned using regular expressions to search for the presence of the diagnosis codes in a patient’s list of diagnosis codes. Patients with serum creatinine values > 1.5 mg/dL were assigned one point, while those with serum creatinine values > 2.0 mg/dL were assigned two points. The nephropathy category was the maximum point assignment of either the diagnosis code or the serum creatinine value. All of these categories were summed and calculated as the maximum for each patient per visit.

An Elixhauser score (0-36) was calculated for each patient using diagnosis codes and assigning a single point for each comorbidity within the 36 categories established in the literature. The CCI score (0-34) was calculated by giving values to comorbidities in one of 17 categories. Patients who had more serious conditions were given more points for these conditions as per CCI guidelines in the literature: 2 (hemiplegia, moderate or severe renal disease, diabetes with end stage-organ damage, tumor without metastasis, leukemia, lymphoma); 3 (moderate or severe liver disease); 6 (metastatic solid tumor, AIDS). Age brackets (<50, 50-60, 60-70, 70-80, >80) were determined by calculating the difference between ICU stay and date of birth and assigned point values of 0, 1, 2, 3, 4, and 5 respectively. The diagnosis score and age bracket score values were then summed and calculated as described above for the DCSI.

**Prediction Modeling**

The outcome of interest was all cause mortality in the index visit. All model fitting was conducted using packages from the Julia programming language. The GLM.jl package was used to fit binomial logistic regression models. Model creation utilized data from 70% of the sample, while the remaining 30% of the sample was used for model validation. Using binomial logistic regression, all of the variables were used in bivariate analyses to correlate with the outcome of mortality. The significant variables with a p-value less than 0.05 were combined for multivariate analysis.

Receiver operating characteristic (ROC) curves were generated by varying the thresholds of each model and the models were compared based on their area under the curve (AUC) values. Each index was first run by itself and compared to one another. Then all of the metrics that had proved significant on preliminary logistic regression analysis were compiled into a final binomial logistic regression model (three indices, demographic variables, and diabetic health metrics).

The DecisionTrees.jl package was used to fit random forest models. All of the variables extracted from the MIMIC-III database were used in this analysis. The following meta-parameters were used: three features randomly selected at each node, 5000 total trees, maximum of five observations in leaf nodes, and a maximum depth of ten. All meta-parameters were validated and selected using five-fold cross validation.

The SciKitLearn.jl package was used for fitting L1 penalized regression (i.e., least absolute shrinkage and selection operator [Lasso]) models on a combined model that contained all of the significant variables from the logistic regression models to determine which variables had the greatest impact on mortality. The following meta-parameters were used: 6 models with varying alpha values from $10^{-4}$ to $10^{0}$.

**Results**

Of the 10,318 patients in the MIMIC-III database with diabetes as a diagnosis, 4111 diabetic patients had values for HbA1c and Blood Glucose; all patients in this subset also had values for Serum Creatinine. Of those patients, 3729 (90.7%) survived, while 382 (9.3%) died during their ICU stay. Summary statistics for this patient population are
shown in Table 1. The six health related variables that were used in the binomial logistic regression are shown at the bottom of Table 2. Differences between the patients with the outcome of death and those who were alive are shown. Of note, there was an increase in all of the health metrics in the subset of patients with a death outcome, other than that of mean A1c value and insulin status. An increase in these categories mirrors an increase in a negative outcome. Patients who died in the ICU on average had 1.71 HbA1c values recorded as opposed to those who survived who had 2.39 HbA1c values. Patients who died in the ICU furthermore had on average 16.23 blood glucose measurements as opposed to those who survived who had 11.54 blood glucose values recorded.

Table 1: Summary statistics on the cohort of patients who died and those who survived

<table>
<thead>
<tr>
<th>Category</th>
<th>Alive</th>
<th>Dead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;50</td>
<td>490</td>
<td>20</td>
</tr>
<tr>
<td>50-60</td>
<td>725</td>
<td>48</td>
</tr>
<tr>
<td>60-70</td>
<td>1023</td>
<td>103</td>
</tr>
<tr>
<td>70-80</td>
<td>926</td>
<td>112</td>
</tr>
<tr>
<td>&gt;80</td>
<td>565</td>
<td>99</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1527</td>
<td>171</td>
</tr>
<tr>
<td>Male</td>
<td>2202</td>
<td>211</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>94</td>
<td>7</td>
</tr>
<tr>
<td>Black</td>
<td>511</td>
<td>69</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>183</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>447</td>
<td>35</td>
</tr>
<tr>
<td>White</td>
<td>2494</td>
<td>261</td>
</tr>
<tr>
<td>Insurance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Government</td>
<td>112</td>
<td>1</td>
</tr>
<tr>
<td>Medicaid</td>
<td>291</td>
<td>23</td>
</tr>
<tr>
<td>Medicare</td>
<td>2238</td>
<td>294</td>
</tr>
<tr>
<td>Private</td>
<td>1061</td>
<td>60</td>
</tr>
<tr>
<td>Self Pay</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>Admission Type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elective</td>
<td>649</td>
<td>15</td>
</tr>
<tr>
<td>Emergency</td>
<td>3015</td>
<td>362</td>
</tr>
<tr>
<td>Urgent</td>
<td>65</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2: Average values for variables of interest in cohort of patients who died and those who survived

<table>
<thead>
<tr>
<th>Variable</th>
<th>Alive Mean</th>
<th>Death Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA1c</td>
<td>7.472 ± 1.86</td>
<td>6.965 ± 1.45</td>
</tr>
<tr>
<td>CCI Score</td>
<td>5.454 ± 2.31</td>
<td>7.113 ± 2.49</td>
</tr>
<tr>
<td>DCSI Score</td>
<td>2.974 ± 1.81</td>
<td>4.166 ± 1.72</td>
</tr>
<tr>
<td>Elixhauser Score</td>
<td>4.449 ± 1.86</td>
<td>5.835 ± 2.00</td>
</tr>
<tr>
<td>Insulin Status</td>
<td>0.968 ± 0.13</td>
<td>0.966 ± 0.18</td>
</tr>
<tr>
<td>Mean Glucose</td>
<td>156.4 ± 48.38</td>
<td>169.07 ± 57.7</td>
</tr>
</tbody>
</table>

Figure 1 shows the ROC curve for the models explored. The AUC for the DCSI only model was 0.694, while the AUC for the Elixhauser model and CCI model alone were 0.682 and 0.656 respectively. When all of the metrics were combined into one binomial logistic regression, the AUC was 0.785 suggesting that combining the three
metrics was the most accurate in predicting the outcome of mortality. Additionally, the AUC for the random forest model was 0.787.

The best performing logistic regression model was the Combined Model that showed the best AUC. Table 3 shows the significance and coefficients for each of the major metrics that contributed to the Combined Model. Each of the metrics showed significance, with the largest coefficient related to Elixhauser Sum and admission type. Ethnicity, insurance, age bracket, gender, and insulin status were not shown to be significant predictors of mortality as they had p-values greater than 0.05.

Table 3: Binomial logistic regression modeling results with coefficients, standard errors, z values, and statistical significance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>Z-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCSI Sum</td>
<td>0.173</td>
<td>0.034</td>
<td>5.011</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Elixhauser Sum</td>
<td>0.203</td>
<td>0.033</td>
<td>6.146</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CCI Sum</td>
<td>0.152</td>
<td>0.027</td>
<td>5.612</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Mean Glucose</td>
<td>0.010</td>
<td>0.001</td>
<td>8.337</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Mean A1c</td>
<td>-0.285</td>
<td>0.046</td>
<td>-6.21</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Admission Type</td>
<td>0.455</td>
<td>0.129</td>
<td>3.533</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Each model showed high sensitivity and specificity at various thresholds, with the Combined logistic regression model and random forest models performing the best. Table 4 shows the sensitivity and specificities after varying the threshold for classifying the linear models and the decision tree algorithm. The combined model showed the highest sensitivity and specificity. The bolded values show the optimal thresholds for sensitivity/specificity combinations. At optimal thresholds of 0.10 for DCSI and Combined; 0.08 for Elixhauser and CCI; and 0.04 for Random Forest, there were 76-78 classified correctly as positive (true positives), 580-777 classified correctly as negative (true negatives), 349-545 classified incorrectly as positive (false positives), and 29-31 classified incorrectly as negative (false negatives). For example, at a threshold of 0.10, the DCSI model classified 78/107 cases correctly as positive, and 674/1126 of the cases correctly as negative. This is in comparison to the CCI model at a threshold of 0.08 that classified 78 correctly as positive and 580 correctly as negative.
Table 4: Sensitivities and specificities of various models

<table>
<thead>
<tr>
<th>Model</th>
<th>DCSI</th>
<th>ELIX</th>
<th>CCI</th>
<th>Combined</th>
<th>Random Forest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold</td>
<td>sens.</td>
<td>spec.</td>
<td>sens.</td>
<td>spec.</td>
<td>sens.</td>
</tr>
<tr>
<td>0.04</td>
<td>97%</td>
<td>9%</td>
<td>97%</td>
<td>15%</td>
<td>97%</td>
</tr>
<tr>
<td>0.06</td>
<td>96%</td>
<td>19%</td>
<td>84%</td>
<td>33%</td>
<td>87%</td>
</tr>
<tr>
<td>0.08</td>
<td>84%</td>
<td>47%</td>
<td>72%</td>
<td>53%</td>
<td>73%</td>
</tr>
<tr>
<td>0.1</td>
<td>73%</td>
<td>60%</td>
<td>54%</td>
<td>73%</td>
<td>62%</td>
</tr>
<tr>
<td>0.12</td>
<td>32%</td>
<td>84%</td>
<td>54%</td>
<td>73%</td>
<td>38%</td>
</tr>
<tr>
<td>0.14</td>
<td>32%</td>
<td>84%</td>
<td>36%</td>
<td>86%</td>
<td>38%</td>
</tr>
<tr>
<td>0.16</td>
<td>24%</td>
<td>92%</td>
<td>36%</td>
<td>86%</td>
<td>25%</td>
</tr>
</tbody>
</table>

A series of Lasso models were used to estimate which variables had the largest impact on the model. Significant variables from the logistic regression models were applied to determine which metrics carried the greatest significance. The results appear in Figure 2. Mean glucose, Elixhauser score, and CCI score were the last three metrics whose coefficients shrank to 0 as alpha increased.

**Figure 2:** Lasso logistic regression models with increase penalty terms (alpha)

Discussion

Mortality in the ICU is a multifactorial issue that depends upon severity of the admitting disease, quality of care, and infections, amongst other issues. Diabetic patients in the ICU present a unique problem of management of a chronic condition while simultaneously managing acute exacerbation of a comorbidity. It is important to predict mortality of diabetic patients on ICU admission because this will allow for better management of care in the ICU. The summary data shown in Table 2 suggest that no single existing comorbidity score or single variable is able to accurately predict mortality, as the means of the individual predictors demonstrate overlap between the subsets of patients who survived and those who died. Thus, the use of multivariate linear modeling and machine learning algorithms are useful to better predict the complex interplay between the variables in the prediction of mortality.

As a first study, mortality was considered a binomial outcome variable. Future work might include additional outcomes (e.g., bins consisting of range of time to death values). Additionally, this study focused on patients who had HbA1cs and Blood Glucose values. The lack of HbA1c data for certain diabetic patients does not seem to be
Based on clinical differences between the two populations, but rather as a function of missing data from a patient’s prior history at the hospital. Future work could investigate the differences between diabetic patients in the MIMIC-III database who had HbA1cs and those who did not. Additional characterization of these two groups will give further insight into the problem at hand.

Based on the results of the logistic regression analysis, the variables that significantly correlated with mortality were mean glucose of each ICU visit, mean HbA1c for each patient, index admission type for each patient, and the various indices’ scores (Table 3). Patient ethnicity, admission location, insurance, and insulin status were not significantly associated with mortality. Lack of correlation with insulin status is likely due to lack of stratification of this variable (with 96.6% of the population using insulin). Insulin use in type 2 diabetics begins in later stages and with more advanced disease, suggesting that this population has had a long diabetic history. A recent study by Vincent, et al., described this phenomenon by showing that insulin treated diabetes did not increase the risk of mortality in the ICU. However this study did not stratify diabetic patients with other variables as this present study has. Perhaps surprisingly, neither insurance nor patient ethnicity were an indicator of future mortality. Previous studies have shown that insurance type has a significant effect on diabetic treatment with 36% of privately insured patients getting all recommended preventative treatment vs. 16% for publicly insured. Despite this result, however, insurance did not prove to affect the model.

In this particular data set, the mean blood glucose was the variable most strongly associated with diabetic mortality in the ICU based on the series of Lasso models (Figure 2). Based on these results, every increase in mean blood glucose by 1 mg/dl led to an 0.010 increase risk in death suggesting that 25 mg/dl changes in blood glucose can increase log odds risk by 25%. This result is of significant interest because the question of how to manage diabetic blood sugar in the ICU is highly debated.

The first randomized control trial of blood glucose in the ICU, the Leuven trial showed this trend, suggesting that tight control to 80-110 mg/dl was more successful in treating patients than the control group which had glucose values of 180–215 mg/dL. Compared to the control group, absolute mortality was decreased by 3%. In contrast, the NICE-SUGAR trial showed that tight control of diabetic blood sugars led to an absolute mortality increase of 3%. Instead, they argued that physicians should target 140 mg/dl for their patients as they had targeted in their control group. The latter article has been taken to be the standard of care leading many ICUs to practice maintaining higher levels of glucose despite the contradicting evidence shown in the earlier studies. Both research studies that were performed used different methods of blood glucose measurement (blood gas analyzers vs. hand glucose monitors). The result from this current study suggests that this issue may need to be revisited as the greatest correlating factor with mortality was that of mean hospital blood glucose level.

The negative effect of HbA1c on the risk of death is problematic suggesting that as a variable there is not a normative distribution. HbA1c is clinically used as a measure of diabetic health over the past three months indicating how well a patient’s blood sugar has been maintained. While mean glucose during the patient’s stay is most likely to affect the immediate situation, a higher HbA1c value is correlated with higher risk of complications by itself. Perhaps the number of HbA1c tests in the patients who died (1.78) versus those who survived (2.24) may have had an impact on this variable’s effect on the model. A future analysis that warrants investigation is whether or not the most recent A1c has as strong of an impact as well as the number of a patient's HbA1cs or Blood Glucoses. Additionally, further investigation needs to be conducted into whether these patients were being managed on stronger medications. It is possible that stronger medication regimens are more strongly correlated to a low HbA1c value than the actual diabetic health of the patient. This type of analysis is not possible with the MIMIC-III database as it only contains data from previously documented visits within the same health system and many patients do not have records other than the current ICU stay. Future studies validating this algorithm should look into the direct effect of HbA1c on ICU mortality.

Amongst the various indices, all three had similar AUC values with the DCSI having a slight advantage. In comparing the three linear models (Table 4 and Figure 1), the AUC for the DCSI model was 0.694 as compared to 0.682 and 0.656 for the Elixhauser model and CCI model respectively. Despite the slightly better performance of the DCSI, none of these three models performed at a level of clinical efficacy. The fact that the DCSI, however, did perform better in detecting diabetic mortality than the other metrics, suggests that its use may be expanded past its previous uses: outpatient, inpatient non-ICU hospitalization, and healthcare cost analysis. Of note, previous studies that validated the DCSI did not compare it to these more accepted comorbidity indices of the CCI and Elixhauser. There is significant overlap between the three indices, with variance in (a) the categorization of the various comorbidities, (b) the inclusion of age as a metric, (c) the inclusion of serum creatinine, and (d) the
granularity of the metric. When the three indices were combined, the AUC of the model became the highest at 0.785 suggesting that this model was the best at predicting mortality in these patients. At a threshold of 0.08, the sensitivity of the model was 80%, and the specificity was 62%, while at 0.10, 73% and 71% respectively. These values suggest that this model is a good predictor of outcomes for diabetic ICU patients.

These AUC values are higher than that reported in the literature of CCI predictability alone in 30-day mortality using retrospective data (0.755). When the model from Stavem et al. was performed upon the dataset, we achieved an AUC of 0.704. The binomial regression and random forest models outperformed this model with an AUC of 0.785 and 0.787 respectively. Additionally, unlike the study performed by Stavem et al., gender was not seen to be predictive in the logistic regression model, as it had a probability > 0.05. One reason for this discrepancy may be due to their choice of outcomes of 30 day and one year mortality, whereas this current study predicted outcomes based on the length of the index ICU visit, which varied. Another potential confounder is that their study was conducted in a nine patient ICU unit at a hospital in Norway and may not be generalizable to data collected in a large tertiary care center in the U.S.

In comparison, the random forest modeling was comparable to the other models in predicting mortality suggesting that it has use in predicting which of these patients will have the greatest risk of mortality. The AUC value of this model was the highest of any of the models at 0.787. At a threshold of 0.06, the model had a sensitivity of 70% and specificity of 73%. This combination suggests that it is a useful method for predicting mortality. Further optimization of the model may allow for an even better prediction value. However, the random forest model still performed slightly better than the logistic regression suggesting it may be better used in this case.

When mean glucose is removed from the combined logistic regression model, the AUC decreases slightly to 0.757, but still outperforms all of the individual indexes alone. This suggests that based on four variables (diagnoses at admission, type of admission, patient’s HbA1c history, and age), one can predict with a sensitivity of 76% and specificity 60% which patients will be likely to die during their stay. Prospective studies are needed to assess whether this relationship is merely correlation or causation. This is in stark contrast to previous studies leveraging machine learning algorithms to predict mortality in ICU patients, including the Super ICU Learner Algorithm as well as the Artificial Neural Networks (ANN).

While both of these studies showed improved AUCs than the present study (0.85), the number of variables needed for prediction rise from five to between 15-35 including various labs, vital signs, and surgical history. Complex models with multiple variables take longer to calculate and require more comprehensive labs and vitals upon admission. Additionally some of the metrics may be heavier on computing power. The model presented in this study, however, requires only a few aspects of a patient’s prior medical history. A case can be made for a simpler, faster to calculate modeling approach for a number of clinical settings. For example, the approach developed here could be used as a pragmatic triage mechanism excluding mean glucose.

A potential limitation of this study is the use of a random 70/30 class sampling for analysis as there was a class imbalance in which less than 10% of the total were positive cases. Future work may consider the use of class balancing techniques such as Synthetic Minority Over-Sampling Technique (SMOTE) to account for this imbalance. It was felt that the imbalance would not significantly invalidate the robustness of the model. Other potential future work may look into whether a patient may be admitted for diabetes care (e.g., with either hypo- or hyperglycemia) have different mortality outcomes compared to those who were admitted primarily with non-diabetes related issues. It should be noted that such an analysis using the MIMIC-III database can be difficult as patients are admitted with a list of ordered diagnosis codes as well as a free-text, non-coded chief complaint. Deciphering the meaning of chief complaints mapped to codes (e.g., ICD-9-CM) through use of natural language processing could therefore involve assumptions. For example, a patient may be admitted for a certain condition, but experience hypo- or hyperglycemia while in the hospital – this would be impossible to distinguish from patients who were admitted primarily for failure to control their diabetes.

Further studies need to be conducted to test other machine learning models against the logistic regression and random forest models explored in this study to see if better prediction values may be obtained. Potential implementation of deep learning techniques, such as ANNs, may also allow for significant improvement on the model. Additionally, this model needs to be tested directly against established algorithms (e.g., APACHE-II, SAPS III, and SOFA) to gauge its performance against the clinical gold standards to see both efficacy and the amount of time it would take to complete both. Finally these models need to be validated in prospective cohorts with analysis on length of stay as a variable. Future work could also examine whether modifications of variables in the predictive model during the ICU stay, alter mortality risk.
Conclusion

Managing diabetic patients in the ICU presents a unique problem of managing an acute emergency while maintaining care for a severe chronic disease that may have an immediate impact on survival. This study shows the utility of machine learning modeling to predict which patients are most likely to die in the ICU and gives the opportunity to better treat critical patients, by utilizing only five readily available variables. The promising results set the foundation for future work in developing rapid and robust classification algorithms that leverage the minimal amount of available data.

Acknowledgements

This work was funded in part by the Brown Center for Biomedical Informatics and National Institutes of Health grant U54GM115677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

The Data Gap in the EHR for Clinical Research Eligibility Screening

Alex Butler, BA1, Wei Wei, PhD1, Chi Yuan, MS1, Tian Kang, MA1, Yuqi Si, MS2, Chunhua Weng, PhD1

1Department of Biomedical Informatics, Columbia University, New York City, New York; 2The University of Texas Health Science Center at Houston, Houston, Texas

Abstract

Much effort has been devoted to leverage EHR data for matching patients into clinical trials. However, EHRs may not contain all important data elements for clinical research eligibility screening. To better design research-friendly EHRs, an important step is to identify data elements frequently used for eligibility screening but not yet available in EHRs. This study fills this knowledge gap. Using the Alzheimer’s disease domain as an example, we performed text mining on the eligibility criteria text in ClinicalTrials.gov to identify frequently used eligibility criteria concepts. We compared them to the EHR data elements of a cohort of Alzheimer’s Disease patients to assess the data gap by using the OMOP Common Data Model to standardize the representations for both criteria concepts and EHR data elements. We identified the most common SNOMED CT concepts used in Alzheimer’s Disease trials, and found 40% of common eligibility criteria concepts were not even defined in the concept space in the EHR dataset for a cohort of Alzheimer’s Disease patients, indicating a significant data gap may impede EHR-based eligibility screening. The results of this study can be useful for designing targeted research data collection forms to help fill the data gap in the EHR.

Introduction

Randomized clinical trials (RCTs) are the well-regarded gold standard for generating high-quality medical evidence1. The success of RCTs depends on successful enrollment1,2, which remains the No.1 barrier to RCTs. According to the recent statistics, only 2-4% of adult patients with cancer participate in RCTs, and this number remained unchanged since 19943,4. Inefficient or unrepresentative participant recruitment can cause study delays, increase costs, weaken the statistical power of analysis, and finally, may lead to failed clinical trials4.

A major bottleneck step in RCT recruitment is eligibility screening5. However, conventional methods for eligibility screening involves laborious manual review of the syntactic rules and semantic concepts in eligibility criteria and clinical data sources5,6. This process is not only time-consuming, but also expensive: the cost of eligibility screening is usually not compensated through contracts supporting CTs, and the expense can go up to $336.48 per participant5.

Much effort7,8 has been made to advance automated identification of eligible patients in the biomedical informatics research community. In the meantime, Electronic Health Record (EHR) data have been recognized as an important clinical data source and were adopted in multiple automated identification methods7,10. EHR-based automated approaches have been reported to reduce workload by up to 90%7 and almost reached the theoretical maximum area under ROC curve6.

A concern of EHR-based eligibility screening is that EHRs may not contain all important data frequently used for eligibility screening since EHRs are designed for patient care rather than clinical research. Our previous study in cancer trial eligibility criteria showed that a lot of eligibility criteria used in cancer trials are not present in EHR data so that clinical research coordinators creatively invented a list of “major eligibility criteria” for patient screening to optimize the efficiency of patient screening11. A recent study by Köpcke et al. showed that on average 55% of eligibility criteria required data elements are present in EHR. However, there are three major limitations of their study: (1) only numeric and structured data elements in EHRs like checkboxes and dropdown menus were included in analyses so that EHR narratives were excluded; (2) EHR data from five participating hospitals were not harmonized using any common data model, resulting in unaccounted overlaps or inconsistency among available EHR data elements across sites; (3) the whole process was manual so that patient characteristics (i.e., clinical entities) were manually identified from free-text eligibility criteria followed by assignment of semantic categories, which were again manually mapped to EHR data elements, making their method not scalable.

This study presented here shares the same goal of Köpcke’s study but contributes a novel scalable data-driven approach by leveraging the public clinical trial summary text and the publicly available synthetic clinical data. Next we will describe our methodology details and results as well as implications.
Methods

To overcome the limitations of Köpcke’s study, we extracted common data elements from free-text eligibility criteria\(^{12,13}\) for Alzheimer’s disease (AD) and represented both EHR data elements and eligibility criteria concepts using The Observational Medical Outcomes Partnership (OMOP)\(^{14}\) Common Data Model (CDM) supported by the Observational Health Data Sciences and Informatics (OHDSI)\(^{15}\) consortium (Figure 1).

The OMOP CDM has been adopted by active scientific consortiums such as OHDSI\(^{15}\) and eMERGE\(^{16}\), and has included about 1.26 billion patients as of October 2017. The OMOP CDM-standardized EHR ensures the semantic interoperability of EHR data from multiple participating sites. The sheer number of patients will allow large sample sizes and likely lead to more generalizable study results. Free-text eligibility criteria were automatically processed using Eligibility Criteria Information Extraction (EliIE)\(^{12}\), an open-source information extraction system for structuring eligibility criteria according to the OMOP CDM, and then extracted information (e.g., clinical entities) was stored in a relational database\(^{13}\). The fully automated eligibility criteria processing techniques make our method highly scalable and improve the efficiency of large-scale studies.

As the first step for methodology illustration, we used eligibility criteria from 1,587 clinical trials for AD and a de-identified EHR dataset, Synthesized Public Use File (SynPUF) 1%, to study the data gap. The publicly available SynPUF 1% dataset, which includes a set of over 116,350 patients’ de-identified EHR structured data points, served as the clinical data source. We mapped clinical entities in eligibility criteria to The Systematized NOMenclature of MEDicine – Clinical Terms (SNOMED CT)\(^{17}\) terms (hereafter referred to as “variables”), merged relevant variables and created a list of unique common variables. SNOMED CT was chosen as the ultimate clinical database in this analysis because it has been preferred as the encoding terminology for clinical concepts by researchers on various other projects.\(^{18}\) We picked a subjective threshold of “being present in at least 15 trials” to select common variables, visualized the relations among the variables and their parents, and analyzed the prevalence of the variables in an EHR dataset. For the purposes of this analysis, we focused on the 19,570 patients who had a previous diagnosis of Alzheimer’s Disease within the SynPUF 1% dataset (hereafter referred to as “the EHR dataset”) as the clinical data source in this study. OHDSI ATLAS, a web-based open source software available at http://www.ohdsi.org/web/atlas for scientific analyses of observational data was adopted to identify qualified patient records from the EHR. The details of are provided below (Figure 2).

**Figure 1.** Overview of the study design for comparing OMOP CDM-based criteria and EHR data for AD trials

**Figure 2.** The eight-step workflow of this study.

**Step 1. Automated Eligibility Criteria Extraction from AD Trials and Concept Standardization**

Free-text eligibility criteria were downloaded from TheClinicalTrials.gov, reformatted using the previously published open-source EliIE\(^ {12}\) system, and stored in a public relational database (https://github.com/Yuqi92/DBMS_EC)\(^ {13}\). All the eligibility criteria of 1,587 AD trials (collected until September 2016) were represented using the OMOP CDM v5.0 model, which allows focusing on four classes of entities: *condition, observation, drug/substance, and procedure*
or device. A total of 9,261 unique clinical entities were identified from all of the eligibility criteria. For analysis, corresponding modifiers (e.g., qualifier, measurement) and inclusion/exclusion status were attached to each entity.

**Step 2. Manual Curation of Clinical Entities**

A manual review of unique clinical entities was performed by a medical student (AB). Modifications were made to produce a simplified list of clinical entities (e.g., AD was used to refer to Alzheimer’s disease). To identify the relevant entities, all entities were sorted alphabetically, so word-similar entity comparison was possible as has been done algorithmically by Varghese and Dugas. All reasons for modification were captured and can serve as evidence in the future for eligibility criteria terminology guidelines.

**Step 3. UMLS Concept Recognition**

The clinical entities in the simplified list were mapped to the Unified Medical Language System (UMLS) Metathesaurus, which was chosen because it is the largest thesaurus in the biomedical domain. This mapping was performed via a widely adopted NLP system developed by the National Library of Medicine, MetaMap. MetaMap was chosen over other NLP systems because of its widespread adoption, easy learning curve and batch request functionality, which allowed large blocks of text to be analyzed simultaneously. For clarity, all phrases contained in the original entity list will be referred to as “entities” and all terms found in the UMLS Metathesaurus will be referred to as “concepts.” The configuration of MetaMap query options were as below:

- `-JSONf 2` (formatted JSON output),
- `-g` (Allow Concept Gaps),
- `-z` (Term Processing),
- `-Q 4` (Composite Phrases),
- `-y` (Use Word Sense Disambiguation),
- `-E` (Indicate Citation End; required for batch scheduler)

Figure 3 illustrates this concept recognition process. When multiple phrases contain one or more concepts in a query, the term with the highest MetaMap score was retrieved. In the case that multiple phrases containing 1 or more concept were returned with identical MetaMap scores, the phrase with the lowest level of clinical specificity was chosen to not exclude any concepts. Review of the simplified entity list found numerous multi-term entities, so single term retrieval was not performed.

**Figure 3.** The process of deriving SNOMED CT terms from clinical trial eligibility criteria. 42,131 clinical entities were extracted from the eligibility criteria of 1,587 clinical trials. A simplified list of 4,260 clinical entities was generated following manual review and filtration, and this list was mapped first to 3,294 UMLS concepts, and then to 1,991 SNOMED CT variables, of which 304 variables occur in more than 1% of all trials (i.e., 15 trials).
Step 4. UMLS CUI Manual Review and Revision

There were a number of data quality issues identified when performing concept extraction. A total of 3,610 manual edits were made to the “master list” for clinical entities as tracked by our computer with the six main types, including typos, plural, trimmed, other formatting reason, simplification, and multi-term (Table 1). Therefore, the identified UMLS concepts and associated Concept Unique Identifiers (CUIs) were manually reviewed and corrected by a medical student (AB). The corrections were performed for two primary reasons: (1) simple corrections which are applied when the CUI of a concept is replaced by a more appropriate CUI, and (2) type corrections which are applied when the CUI of a concept is replaced by a CUI of a more appropriate type according to UMLS coding.

Step 5. Mapping to SNOMED CT

For every UMLS concept, its corresponding term in SNOMED CT was identified. Due to the design of UMLS Metathesaurus as a hub for numerous terminologies, the SNOMED CT variables associated with the UMLS concepts were used when such variables were possible. In the case that no SNOMED CT variable was found, a manual search of the SNOMED CT terminology was conducted to identify the closest available match (Figure 3). Manual modifications were also performed for SNOMED CT types which were inappropriate for use in eligibility screening. For example, “alanine aminotransferase (substance)” was changed to “alanine aminotransferase measurement (procedure).”

Table 1. Manual revision of clinical entities.

<table>
<thead>
<tr>
<th>Types of Revision</th>
<th>Example</th>
<th>Times</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formatting; Typo</td>
<td>delerium &lt;-&gt; delirium</td>
<td>207</td>
</tr>
<tr>
<td>Formatting; Plural</td>
<td>cancers &lt;-&gt; cancer</td>
<td>253</td>
</tr>
<tr>
<td>Formatting; removal of non-informative words</td>
<td>heart rate measurement &lt;-&gt; heart rate</td>
<td>364</td>
</tr>
<tr>
<td>Formatting; removal of abbreviations</td>
<td>absolute neutrophil count (ANC) &lt;-&gt; absolute neutrophil count</td>
<td>1768</td>
</tr>
<tr>
<td>Simplification</td>
<td>asthmatic conditions &lt;-&gt; asthma</td>
<td>573</td>
</tr>
<tr>
<td>Breaking down long phrases to logically-connected single phrases</td>
<td>basal or squamous cell carcinoma &lt;-&gt; basal cell carcinoma or squamous cell carcinoma</td>
<td>445</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>3610</strong></td>
</tr>
</tbody>
</table>

Step 6. Establishing a “Master List”

Trial occurrences were tracked for each clinical entity and carried through to mapped SNOMED CT variables to calculate an overall trial frequency. SNOMED CT variables chosen for the “master list” were found in at least 1% of all trials, meaning they were used as an eligibility criterion in at least 15 trials.

Step 7. Visualization of Selected SNOMED CT Variables

Since SNOMED CT maintains a hierarchical structure, the parents of all variables present in the “master list” were captured. All of the “master list” variables, their parent variables, and the “is-a” hierarchical relations were stored in JSON files and visualized using a modified d3js package. Also, the trial frequency for each variable was also obtained and stored within the corresponding JSON file. Of note, every parent of a “master list” variable was considered to have the same trial frequency as its child.

Step 8. Assessment of SNOMED CT Variable Coverage in the EHR Dataset

The SNOMED CT ID associated with each SNOMED CT variable in the “master list” was queried in ATLAS and the record count (RC) and descendant record count (DRC) were returned. RC indicates the number of times a specific variable is found in the EHR dataset, and DRC indicates the number of times a specific variable and its descendants are found in the dataset. SNOMED CT variables were further classified into five sets:

1. categorical variables (e.g., the presence of Parkinson’s Disease) that are available in EHR
2. continuous variables (e.g., age) that are available in EHR
3. variable not found in EHR, but can be derived from the existing EHR variable, such as “chronological age” can be derived from variable “date of birth”
(4) variables not available in EHR, but the data could be collected from a patient without medical training, such as questions in Mini-Mental Status Exam
(5) variables not available in EHR, and the information could not be provided by a patient without medical training, such as “General Metabolic Function”
(6) variables not found in EHR, and not relevant for eligibility screening, such as “Psychiatric”

Results

The 42,131 entities identified in clinical trial eligibility criteria contained 9,261 unique entities, 1,930 of which corresponded to medication information which were not included in this analysis. Manual review of the remaining 7,331 unique non-medication entities simplified the list to 4,260 entities. To reach this simplified list, 3,610 manual changes were made. 2,591 changes were made for formatting reasons (e.g. AD, AD Disease -> Alzheimer’s Disease), 574 changes were made for simplification reasons (e.g. asthmatic conditions, adult asthma -> asthma) and 445 changes were made for ‘multi-term’ entities (e.g. basal or squamous cell carcinoma -> basal cell carcinoma or squamous cell carcinoma). A total of 4,260 unique clinical concepts were mapped to UMLS concepts via MetaMap, resulting in 4,026 unique MetaMap term sets (e.g. basal cell carcinoma or squamous cell carcinoma is a single ‘term set’ as the phrase was extracted from an eligibility criterion, but each underlined section is handled as a separate UMLS concept). A total of 111 manual searches were performed, including 66 searches for multi-term clinical entities, one for a typo in the entity, and 44 for inaccurate MetaMap mapping as assessed by the medical student (AB). After sorting, the final UMLS concept list was composed of 3,294 unique concepts. Of note, it was observed on manual review that many of the lab tests being used for eligibility assessment were found to be of UMLS type “Amino Acid, Peptide, or Protein” so all concepts of this type were re-queried searching only for concepts with the type “Laboratory Procedure” or “Laboratory or Test Result”.

Direct matching to SNOMED CT using the UMLS Metathesaurus returned 1,991 unique SNOMED CT variables (e.g. basal cell carcinoma [UMLS code C00071117] is directly linked to epithelioma basal cell [SNOCT code 275265005] within databases). 56 variables were manually added by the direct query in the SNOMED CT Browser as no direct UMLS to SNOMED CT connection existed. Further, during the manual review, it was observed that some UMLS concepts which had no direct SNOMED CT equivalent could be applicable to a SNOMED CT variable returned for another concept, so the trial count and additional information was attached from both concepts to the single SNOMED CT variable. When filtered by variables identified in at least 15 trials out of the entire list, a “master list” was generated containing 318 UMLS concepts and 304 SNOMED CT variables (14 concepts had no correlated SNOMED CT variable). The UMLS concepts found in the “master list” were found in 1491 of the 1512 queried trials, i.e., a trial coverage of 98.6%.

Visualization of The Common Eligibility Criteria SNOMED CT Variables and their hierarchical relations

The highly prevalent eligibility criteria concepts in AD trials are listed in Table 2. Since there exist hierarchical relations among these concepts, an online visualization was also generated for these concepts. Each node in the visualization is a common eligibility criteria concept in AD trials followed by its prevalence. For example, “mental disorder” is a node with prevalence of 82.21% because it is used by 82.21% of AD trials for patient screening. The visualization of “master list” concepts and their super classes can be observed at http://htmlpreview.github.io/?https://github.com/Butler925/Alz_viz/blob/master/index_git.htm.

Table 2. The most commonly adopted eligibility criteria variables and their prevalence in AD trials (the last column with column header as “#” indicates the number of parent concepts)
**Clinical history and observation findings**

<table>
<thead>
<tr>
<th>SNOMED-CT Semantic Type</th>
<th>Trial Count</th>
<th>Prevalence in Trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>250171008</td>
<td>64.55%</td>
<td>finding</td>
</tr>
<tr>
<td>26929004</td>
<td>64.29%</td>
<td>disorder</td>
</tr>
<tr>
<td>254291000</td>
<td>60.65%</td>
<td>staging scale</td>
</tr>
<tr>
<td>273249006</td>
<td>60.65%</td>
<td>assessment scale</td>
</tr>
<tr>
<td>71388002</td>
<td>58.33%</td>
<td>procedure</td>
</tr>
<tr>
<td>363787002</td>
<td>51.19%</td>
<td>observable entity</td>
</tr>
<tr>
<td>273617000</td>
<td>46.63%</td>
<td>assessment scale</td>
</tr>
<tr>
<td>362981000</td>
<td>45.24%</td>
<td>qualifier value</td>
</tr>
<tr>
<td>118222006</td>
<td>41.14%</td>
<td>finding</td>
</tr>
<tr>
<td>12348006</td>
<td>39.62%</td>
<td>disorder</td>
</tr>
<tr>
<td>49601007</td>
<td>39.55%</td>
<td>disorder</td>
</tr>
<tr>
<td>116367006</td>
<td>38.96%</td>
<td>finding</td>
</tr>
<tr>
<td>384821006</td>
<td>38.96%</td>
<td>finding</td>
</tr>
<tr>
<td>118940003</td>
<td>35.78%</td>
<td>disorder</td>
</tr>
<tr>
<td>424144002</td>
<td>34.06%</td>
<td>observable entity</td>
</tr>
<tr>
<td>105727008</td>
<td>34.06%</td>
<td>observable entity</td>
</tr>
<tr>
<td>27550009</td>
<td>33.33%</td>
<td>disorder</td>
</tr>
<tr>
<td>386053000</td>
<td>33.33%</td>
<td>procedure</td>
</tr>
<tr>
<td>362965005</td>
<td>32.41%</td>
<td>disorder</td>
</tr>
<tr>
<td>62914000</td>
<td>32.28%</td>
<td>disorder</td>
</tr>
<tr>
<td>113091000</td>
<td>31.88%</td>
<td>procedure</td>
</tr>
<tr>
<td>123940008</td>
<td>30.16%</td>
<td>disorder</td>
</tr>
<tr>
<td>128927009</td>
<td>28.39%</td>
<td>procedure</td>
</tr>
<tr>
<td>46206005</td>
<td>28.39%</td>
<td>disorder</td>
</tr>
<tr>
<td>66214007</td>
<td>25.66%</td>
<td>disorder</td>
</tr>
<tr>
<td>272099008</td>
<td>24.80%</td>
<td>qualifier value</td>
</tr>
<tr>
<td>230690007</td>
<td>24.54%</td>
<td>disorder</td>
</tr>
<tr>
<td>284061009</td>
<td>23.94%</td>
<td>assessment scale</td>
</tr>
<tr>
<td>56019007</td>
<td>23.35%</td>
<td>finding</td>
</tr>
<tr>
<td>82832008</td>
<td>22.49%</td>
<td>finding</td>
</tr>
<tr>
<td>386086002</td>
<td>21.43%</td>
<td>finding</td>
</tr>
<tr>
<td>230226000</td>
<td>21.16%</td>
<td>disorder</td>
</tr>
<tr>
<td>60342002</td>
<td>21.16%</td>
<td>disorder</td>
</tr>
<tr>
<td>76349003</td>
<td>21.16%</td>
<td>disorder</td>
</tr>
<tr>
<td>118940005</td>
<td>21.03%</td>
<td>disorder</td>
</tr>
<tr>
<td>35489007</td>
<td>21.03%</td>
<td>disorder</td>
</tr>
</tbody>
</table>

**SNOMED CT Variable Assessment**

Overall, the “master list” contained 21 SNOMED CT semantic types and 13 of the 19 highest-level SNOMED CT variable types. The prevalence of these concepts in AD trials is shown in Table 3, with the top 20 shown in Table 4. Of note, the majority of the variables in Table 4 are specific except for variable “Disease”, which is very vague. The less vague but still non-specific example variables are “Systematic Disease” and “History of clinical finding in subject”.

**Table 3.** The counts of trials containing each SNOMED CT semantic type.
<table>
<thead>
<tr>
<th>Disorder</th>
<th>1425</th>
<th>94.25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding</td>
<td>1072</td>
<td>70.90%</td>
</tr>
<tr>
<td>Assessment scale</td>
<td>917</td>
<td>60.65%</td>
</tr>
<tr>
<td>Staging scale</td>
<td>917</td>
<td>60.65%</td>
</tr>
<tr>
<td>Procedure</td>
<td>882</td>
<td>58.33%</td>
</tr>
<tr>
<td>Observable entity</td>
<td>774</td>
<td>51.19%</td>
</tr>
<tr>
<td>Qualifier value</td>
<td>684</td>
<td>45.24%</td>
</tr>
<tr>
<td>Situation</td>
<td>250</td>
<td>16.53%</td>
</tr>
<tr>
<td>Physical object</td>
<td>231</td>
<td>15.28%</td>
</tr>
<tr>
<td>Attribute</td>
<td>163</td>
<td>10.78%</td>
</tr>
<tr>
<td>Linkage concept</td>
<td>163</td>
<td>10.78%</td>
</tr>
<tr>
<td>Body structure</td>
<td>154</td>
<td>10.19%</td>
</tr>
<tr>
<td>Metadata</td>
<td>105</td>
<td>6.94%</td>
</tr>
<tr>
<td>Morphologic abnormality</td>
<td>125</td>
<td>8.27%</td>
</tr>
<tr>
<td>Mother</td>
<td>56</td>
<td>3.70%</td>
</tr>
<tr>
<td>Substance</td>
<td>21</td>
<td>1.39%</td>
</tr>
<tr>
<td>Regime/therapy</td>
<td>33</td>
<td>2.18%</td>
</tr>
<tr>
<td>Environment</td>
<td>19</td>
<td>1.26%</td>
</tr>
<tr>
<td>Environment / location</td>
<td>19</td>
<td>1.26%</td>
</tr>
<tr>
<td>Event</td>
<td>17</td>
<td>1.12%</td>
</tr>
<tr>
<td>Organism</td>
<td>15</td>
<td>0.99%</td>
</tr>
</tbody>
</table>

The top 20 common SNOMED CT terms in AD trials and their prevalence in EHR dataset.

<table>
<thead>
<tr>
<th>SNOMED CT Term</th>
<th>SNOMED-CT ID</th>
<th>Trial Count</th>
<th>Prevalence in Trials</th>
<th>Count of uses in EHR data for AD patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's disease</td>
<td>26929004</td>
<td>972</td>
<td>64.29%</td>
<td>30,262</td>
</tr>
<tr>
<td>Mini-mental state examination</td>
<td>273617000</td>
<td>705</td>
<td>46.63%</td>
<td>0</td>
</tr>
<tr>
<td>Presenile dementia</td>
<td>12348006</td>
<td>599</td>
<td>39.62%</td>
<td>7,089</td>
</tr>
<tr>
<td>Disease</td>
<td>64572001</td>
<td>555</td>
<td>36.71%</td>
<td>12,029,900</td>
</tr>
<tr>
<td>Current chronological age</td>
<td>424144002</td>
<td>515</td>
<td>34.06%</td>
<td>0</td>
</tr>
<tr>
<td>Mental disorder</td>
<td>74732009</td>
<td>499</td>
<td>33.00%</td>
<td>505,870</td>
</tr>
<tr>
<td>Magnetic resonance imaging</td>
<td>113091000</td>
<td>482</td>
<td>31.88%</td>
<td>63,171</td>
</tr>
<tr>
<td>Cerebrovascular accident</td>
<td>230690007</td>
<td>371</td>
<td>24.54%</td>
<td>4</td>
</tr>
<tr>
<td>Global assessment of functioning - 1993 Diagnostic and Statistical Manual of Mental Disorders- ver.4th</td>
<td>284061009</td>
<td>361</td>
<td>23.88%</td>
<td>0</td>
</tr>
<tr>
<td>Systemic disease</td>
<td>56019007</td>
<td>353</td>
<td>23.35%</td>
<td>0</td>
</tr>
<tr>
<td>Disorder of nervous system</td>
<td>118940003</td>
<td>335</td>
<td>22.16%</td>
<td>780,478</td>
</tr>
<tr>
<td>Substance abuse</td>
<td>66214007</td>
<td>279</td>
<td>18.45%</td>
<td>9,466</td>
</tr>
<tr>
<td>Parkinson's disease</td>
<td>49049000</td>
<td>275</td>
<td>18.19%</td>
<td>0</td>
</tr>
<tr>
<td>Impaired cognition</td>
<td>386806002</td>
<td>260</td>
<td>17.20%</td>
<td>13,375</td>
</tr>
<tr>
<td>Seizure disorder</td>
<td>128613002</td>
<td>240</td>
<td>15.87%</td>
<td>28,586</td>
</tr>
<tr>
<td>Hypersensitivity reaction</td>
<td>421961002</td>
<td>218</td>
<td>14.42%</td>
<td>4,686</td>
</tr>
<tr>
<td>Schizophrenic disorders</td>
<td>191526005</td>
<td>216</td>
<td>14.29%</td>
<td>40777</td>
</tr>
<tr>
<td>History of clinical finding in subject</td>
<td>417662000</td>
<td>207</td>
<td>13.69%</td>
<td>189,543</td>
</tr>
<tr>
<td>Risk identification: childbearing family</td>
<td>386414004</td>
<td>205</td>
<td>13.56%</td>
<td>0</td>
</tr>
<tr>
<td>Clinical dementia rating scale</td>
<td>273367002</td>
<td>204</td>
<td>13.49%</td>
<td>0</td>
</tr>
</tbody>
</table>

The Data Gap

Table 5 shows the counts of SNOMED CT variables from the “master list” for each of the five categories. 60% of the variables from the “master list” were found in the EHR dataset, but data for about 40% of the variables that are not available in EHR could be provided by patients without clinicians’ assessment. Determining if patients could answer some of the criteria that have no data in the EHR largely relied on health literacy and access to their medical records.
Criteria that are considered symptoms or based on clinical discretion (e.g. amyloid deposition, neurological deficit, psychotic symptom) are unanswerable by patients. Further, specific lab test results (e.g. Cobalamin deficiency, laboratory test abnormal) are also considered to be unanswerable by patients as they may not have the health literacy to address these criteria. Those criteria which are considered answerable by patient are broken into three categories: (1) discrete diagnosis (e.g. Parkinson’s Disease, Multiple Sclerosis, Carcinoma of Prostate), (2) answerable with online test (e.g. visual acuity, auditory acuity, memory function), and (3) answerable with structured questions (e.g. Clinical Dementia Rating Scale, Hachinski Ischemia Score, Geriatric Depression Scale). The ‘master list’ with EHR record counts, descendant record counts, and characterization about how a patient can address the criterion is at https://docs.google.com/spreadsheets/d/1R6_xc_iEq34YUWuJLzT26J1kskEGlGmoQCOgrUJiB3w/edit?usp=sharing.

**Table 5.** The count of SNOMED CT variables from the “master list” in the five categories.

<table>
<thead>
<tr>
<th>Category Description</th>
<th>Example</th>
<th>Categories</th>
<th>Total Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>In EHR, categorical</td>
<td>Presenile Dementia</td>
<td>132</td>
<td>181 (60%)</td>
</tr>
<tr>
<td>In EHR, continuous</td>
<td>Laboratory Test</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Not in EHR, can be</td>
<td>Chronological Age</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Not in EHR, answerable by patient</td>
<td>Questions from Mini-Mental Status Exam</td>
<td>59</td>
<td>123 (40%)</td>
</tr>
<tr>
<td>Not in EHR, not answerable by patient</td>
<td>General Metabolic Function</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Not applicable</td>
<td>Psychiatric</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

**Discussion**

*The EHR data gap for eligibility screening*

From Table 4 we can see that multiple variables used frequently for eligibility screening were not present in the EHR, including *mini-mental state exam questions’ answers, global assessment of function, systematic disease, risk identification: child bearing family status, and clinical dementia rating scale*. Rating scales used frequently by researchers are usually not available in EHR dataset but constitute important eligibility criteria concepts for AD trials’ eligibility criteria. Our study showed that 60.65% of AD trials include assessment scales and 1.79% of AD trials include symptom ratings, whose corresponding data are not available in EHRs.

Overall, forty percent of the “master list” SNOMED CT variables could not be found in the corresponding structured EHR dataset for patients with AD. The percentage is comparable with the 55% coverage of patients’ characteristics from the study of Köpcke et al. The two studies’ results suggest fully automated EHR-based eligibility screening may still be impossible with the current schema due to the significant data gap, even though both eligibility criteria and EHR data are well represented using a common data model. An improved model may include patient-reported data in areas where criteria are not available in the EHR to allow for comprehensive eligibility criteria coverage.

*Patient self-reported data as a new data source*

An interesting finding is that 19% of the “master list” SNOMED CT variables did not exist in the EHR but could be answered by patients. The finding suggests the involvement of patients in the eligibility screening process may help recruiting more eligible patients. Successful stories include one by Williams et al. who developed and implemented a computer-assisted interview system in an urban rheumatology clinic, and another by Goncalves et al. who showed that use of patient-facing web forms could capture structured data. However, different opinions also exist. For example, one study by Wuerdeman et al. concluded that patient-reported data are likely not as complete or accurate as the information provided by a provider. Some other barriers also have been reported, such as technological fluency, privacy concerns, and lack of technology infrastructure. Further, given that Alzheimer’s Disease affects a patient’s cognition and often presents in the elderly, this could impact the reliability of patient-reported information so it is important that patient-facing tools would include family members and other stakeholders.

*Reusable variables*

Since the 304 UMLS concepts from “master list” variables were found in 98.6% of all the Alzheimer’s disease clinical trials, the clinical entities associated with these concepts could be adopted as common data elements (CDEs) and may help reducing the workload of future Alzheimer’s disease clinical trials by avoiding assessing some of the 9,261 unique clinical entities. There is no currently established CDE for Alzheimer’s Disease, so the results of this study could serve as an important first step.

*Major Eligibility Criteria*
A similar approach to determine the most relevant eligibility criteria was undertaken by using an interview-style approach. Paulson & Weng highlighted the importance of identifying major criteria in creating an optimal clinical trials recruitment tool. Providing equal weight to each eligibility criterion does a disservice in requiring excessive resources for a diminishing return in screening power, so focusing on those most frequent or more important criteria that allow for more robust eligibility screening provides a very strong advantage.

**Limitations**

This study has multiple limitations. First, only Alzheimer’s Disease clinical trials and SNOMED CT variables were included in this study, and this may result in bias in the coverage estimation. If more diseases and all terminologies from OMOP CDM model were included, the assessment of the information gap between EHR and eligibility criteria would be more accurate. Second, we identified a few discrepancies in our SynPUF dataset which may have impacted our results. For example, Parkinson’s Disease as referenced in the SNOMED CT database found no record counts in patient records, however the dataset used in this analysis identified overlap of Parkinson’s Disease in our dataset when searched outside of the SNOMED database. It is possible that there is a coding issue with our dataset, but the more likely scenario is that Parkinson’s Disease is primarily codified using a different clinical database. Future analyses into data source heterogeneity should also be conducted in an attempt to simplify and centralize how all of this data is referenced. Third, variables such as Cerebrovascular accident requires semantic inference and cannot be aligned literally because EHR data may contain specific incidents of Cerebrovascular accident, not this generic concept. Our current simple approach for aligning concepts in criteria and EHR data was unfortunately unable to find its counterpart in the EHR dataset. One implication of this finding is that we need more sophisticated methods for concept matching that is based on semantic alignment between terms, not just based on term matching. Alternative NLP systems to MetaMap, including MedLEE and cTAKES among others, have shown improved identification of clinical terms and may be used in the future to improve on the results elucidated here.

Lastly, one of the most significant limitations in this study involves the intensive manual review necessary to produce these results and its impact on scalability. As evidenced by the 3,610 manual changes made to the original term list in additional to subsequent type modifications and proof-reading, there is a high level of heterogeneity in clinical terminology found in clinical trial eligibility criteria. This heterogeneity increases the workload associated with performing analyses like this and reduces the confidence in the ultimate results. Further, it reduces the scalability of the methods used here. However, tracking of these manual changes does provide some insight into how to address this heterogeneity. Two of the three most common causes for manual modification, formatting and multiple terms, could be easily addressed by using standard term sets or CDEs as mentioned previously. Standardized lists of terms to be used in Alzheimer’s Disease eligibility criteria would avoid any variation in terms based on formatting discrepancies and would allow for simple handling of multiple term concepts (e.g. could identify basal cell carcinoma and squamous cell carcinoma is both terms existed in a standard list). Manual modifications due to simplification were performed primarily for the simplicity of this analysis, so future studies into addressing term heterogeneity should also focus on this reason for modification.

**Conclusions**

We found 40% of the most commonly used criteria variables in Alzheimer’s trial are not available in the concept space in EHR of the patients with Alzheimer’s disease. The result suggests that EHR-based eligibility screening may not achieve perfect performance due to the information gap. To overcome this challenge, a possible solution could be asking patients for missing information during recruitment when using EHR data for trial-eligible patient screening.

**Acknowledgements**

This research is supported by grant 2R01LM009886-08A1 (PI: Chunhua Weng). Author AB is also supported by grant 5T35HL007616-37 (PI: Rudolph L. Leibel).

**References**


27. Liu LS, Shih PC, Hayes GR. Barriers to the adoption and use of personal health record systems. Proceedings of the 2011 iConference; 2011; Seattle, Washington, USA.


Prototype of a Standards-Based EHR and Genetic Test Reporting Tool Coupled with HL7-Compliant Infobuttons

Jacob K. Crump, BS1, Guilherme Del Fiol, MD PHD1, Marc S. Williams, MD2, Robert R. Freimuth, PHD3
1Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah
2Genomic Medicine Institute, Geisinger, Danville, Pennsylvania, 3Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota

Abstract
Integration of genetic information is becoming increasingly important in clinical practice. However, genetic information is often ambiguous and difficult to understand, and clinicians have reported low self-efficacy in integrating genetics into their care routines. The Health Level Seven (HL7) Infobutton standard helps to integrate online knowledge resources within Electronic Health Records (EHRs) and is required for EHR certification in the US. We implemented a prototype of a standards-based genetic reporting application coupled with infobuttons leveraging the Infobutton and Fast Healthcare Interoperability Resources (FHIR) Standards. Infobutton capabilities were provided by OpenInfobutton, an open source package compliant with the HL7 Infobutton Standard. The resulting prototype demonstrates how standards-based reporting of genetic results, coupled with curated knowledge resources, can provide dynamic access to clinical knowledge on demand at the point of care. The proposed functionality can be enabled within any EHR system that has been certified through the US Meaningful Use program.

Introduction
Access to genetic information in medical practice is becoming increasingly prevalent1. Since the end of the Human Genome Project in 2003, over 1000 genetic tests have become available2, allowing clinicians to more easily obtain an individual’s genetic information3. Methods in informatics are permitting these vast amounts of data to be integrated into clinically relevant systems4. Once discovered and translated into clinical practice, biomarkers in the genome, such as an existing gene or variant, can provide clinical help in diagnoses, predictive medicine, and drug treatment (i.e. pharmacogenomics)5. Pharmacogenomics research, as one example, has been found in some cases to reduce healthcare cost6 and improve patient care7. Pharmacogenomics knowledge is rapidly increasing, but the assimilation of this and other kinds of genetic information into clinical practice has been slow8.

In medical practice, questions that clinicians pose often remain unanswered due to lack of time to search available knowledge resources and lack of self-efficacy in choosing and searching multiple potentially relevant resources9. Unanswered questions associated with knowledge gaps can lead to errors and lower quality care10. Clinically actionable genetic information, in particular, can be difficult to obtain. Genetic knowledge resources are especially needed because clinicians report poor knowledge in genetics and low self-efficacy integrating genetics into their decision-making process8. Therefore, care providers may benefit from access to clinically actionable recommendations to guide them in the integration of genetics into routine patient care decisions. Several online resources exist to provide healthcare professionals and patients genetic information, such as ClinVar, Genetic Testing Resources (GTR), and GeneReviews11,12. However, these resources were designed to serve multiple purposes and audiences, with clinically actionable genetic information being dispersed within and across resources5.

The ClinGen (Clinical Genome Resource) project aims to enable “an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research”13. ClinGen seeks to produce open resources by means of community collaboration to continuously improve understanding of genetic variation in a standardized fashion14. However, even with genetic standards, genetic information needs to be easily accessible within a clinical setting. Therefore, the ClinGen Electronic Health Record Working Group (EHR WG) was tasked to integrate available genomic resources (including ClinGen resources) into electronic health record (EHR) systems via clinical decision support (CDS) tools such as “infobuttons”15,16. Based on the context of the interaction between a clinician and the EHR, infobuttons anticipate clinicians’ information needs and provide automated links to relevant resources. Infobuttons are typically placed on the EHR screen next to data such as problems, medications, and laboratory tests17. A recent systematic review has reported the effect of infobuttons on clinicians, particularly on increased search efficiency, decision enhancement, and learning17. Infobuttons can be implemented in a standards-based fashion
through the Health Level Seven (HL7) Infobutton Standard, which is required for EHR certification in the US Meaningful Use Program\textsuperscript{18}. In a previous study, Heale et al. enabled infobutton access to several genetic information resources through OpenInfobutton, an open source, HL7-compliant software suite designed to help healthcare organizations enable infobuttons within their EHR systems\textsuperscript{19}. In the present study, we aimed to implement HL7-compliant infobuttons within a prototype EHR coupled with a standards-based genetic reporting application to demonstrate how this kind of CDS tool could be used to help provide clinicians with seamless access to genetic knowledge resources.

**Methods**

To accomplish the goals of this study, we leveraged an information display of a multi-gene sequencing panel report that was designed and evaluated in a previous study\textsuperscript{20}. We developed a user interface to represent a prototypical EHR coupled with the multi-gene sequencing panel report and incorporated infobutton features to help clinicians understand other genetic-related information, such as diseases and medications (i.e., pharmacogenomics). The prototype was developed in three steps: (1) configuration of access to genetic resources within OpenInfobutton; (2) development of the prototype EHR user interface with the gene sequencing panel reports coupled with infobutton functionality; and (3) population of the prototype EHR with data from sample laboratory tests and the EHR prototype with mock patient data, respectively. The patient data were represented using HL7 Fast Healthcare Interoperability Resources (FHIR). Infobutton functionality, compliant with the HL7 Infobutton Standard, was implemented with OpenInfobutton. Brief summaries of the multi-gene sequencing panel report, the HL7 Infobutton Standard, OpenInfobutton, and FHIR are provided below.

**Multi-gene sequencing panel report design.** In a study done by Cutting et al., a multi-gene sequencing panel report was iteratively designed and evaluated through clinician feedback on suitable genetic reporting\textsuperscript{20}. The report, known as Genomic Medicine Assistant (GMA), featured the clinically significant gene and variant names discovered in the panel with infobuttons next to those names. The report also featured the clinical significance of the findings, recommended actions, and other relevant features. Clinicians in the study rated the usability of the GMA and the results were favorable.

**HL7 Infobutton Standard.** EHR certification under the US “Meaningful Use” program requires the implementation of infobutton functionality compliant with the HL7 Context-Aware Knowledge Retrieval (“Infobutton”) Standard\textsuperscript{21, 22}. The infobutton standard defines a means to integrate EHR systems with health knowledge resources\textsuperscript{22}. Infobuttons utilize the context of the EHR to anticipate clinicians’ and patients’ information needs; select relevant Web resources; and automatically retrieve relevant information from the selected resources to address those needs. The Infobutton Standard allows EHRs to capture and send context in a standard context information model according to four dimensions: the patient (e.g., age, gender, conditions), the infobutton user (e.g., clinician, patient), the care setting (e.g., inpatient, outpatient), and the EHR task (e.g., order entry, medication list review, laboratory test result review, problem list entry)\textsuperscript{19}. In addition, the standard allows capturing one or more concepts that the user might be interested in, such as a specific medication, gene, or laboratory test.

**OpenInfobutton (www.OpenInfobutton.org).** Infobutton functionality can be implemented through a Web service called an “infobutton manager”, which serves as a broker between EHR systems and multiple knowledge resources. All configuration of infobutton behavior is done within the infobutton manager’s knowledge base. OpenInfobutton is an HL7 compliant, open source software suite that includes an infobutton manager and a Web-based infobutton configuration application called LITE (Librarian Infobutton Tailoring Environment). LITE allows knowledge resource experts, such as medical librarians, to collaboratively configure access to knowledge resources within the infobutton manager’s knowledge base\textsuperscript{23, 24}. Knowledge resource configuration is stored in XML files called knowledge resource profiles, which include the resource’s base URL, the contexts in which the resource is relevant, and the search parameters and terminologies that the resource’s Web service application program interface (API) supports. For resources that are not HL7-compliant, the resource profile also includes mappings between search parameters in the resource’s API and the HL7 Infobutton Standard. Once a resource profile is configured in LITE, it can be made available for download to all institutions using OpenInfobutton via LITE’s resource profile store.

**FHIR (www.hl7.org/fhir/).** HL7 FHIR is an emerging standard that is rapidly gaining attention and adoption by the health IT industry and health care organizations\textsuperscript{25}. The basic building block of FHIR is a resource, which consists of a set of data attributes that are defined based on FHIR data types and bound to standard terminologies whenever applicable\textsuperscript{26}. A few examples of FHIR resources include Patient, Condition, and Observation\textsuperscript{27}. FHIR profiles are available to define the semantic structure of resources for specific use cases through constraints and extensions\textsuperscript{28}. 
Configuration of ClinGen and other genetic resources in OpenInfobutton
A set of clinically useful genetic resources was chosen by consensus among ClinGen EHR WG members. Also with input from the ClinGen EHR WG, resource profiles were selected and fine-tuned from those initially developed by Heale et al. In addition, new profiles were created in LITE for APIs that became available after the work done by Heale et al. All knowledge resource profiles were validated using LITE’s testing tool.

Development of a prototype EHR with gene sequencing panel reports coupled with infobuttons
The goal of our infobutton-enabled prototypical EHR was to demonstrate how genetic knowledge resources can be integrated within EHR systems to help clinicians’ meet their genetic information needs. The prototype user interface was developed in Javascript, HTML, and CSS. Using the same approach, gene sequencing panel reports based on the design by Cutting et al. were included in the prototype EHR. Infobutton “i” icons were placed next to each clinical data element in both the EHR and embedded panel reports. Infobutton icons contained hyperlinks with HL7-compliant OpenInfobutton requests.

Population of the prototypes with data from sample laboratory tests and FHIR compliant JSON files
The gene panel prototype was populated with data from three case vignettes developed by Cutting et al. and sample laboratory test reports from Integrated Genetics. To demonstrate the use of the FHIR standard, data in the EHR prototype were represented according to FHIR resources. Example FHIR resources were selected from two openly available FHIR servers: HAPI (hapifhir.io) and Health Services Platform Consortium (HSPC) Sandbox (sandbox.hspconsortium.org). Using these sample resources as guidelines, FHIR compliant JSON files with FHIR resources were created for three sample patients. These FHIR compliant JSON files were utilized to populate the prototype EHR.

Results
Configuration of ClinGen and other genetic resources in OpenInfobutton
Eight genetic knowledge resources selected by members of the ClinGen EHR Work Group were configured for EHR integration using LITE (Table 1). Although only one (PharmGKB) out of the eight genetic knowledge resources was compliant with the HL7 Infobutton Standard, we were able to successfully connect with all eight resources. Profiles were created for each resource by setting the base URLs, EHR contexts, support for standard terminologies, and mappings between resource-proprietary parameters and the HL7 Infobutton Standard (for resources that were not HL7-compliant). The infobutton user was selected as “Provider” for each resource.

Table 1. Summary of the knowledge resources connected through the HL7 Infobutton Standard via LITE.

<table>
<thead>
<tr>
<th>Infobutton Resource</th>
<th>Information Available</th>
<th>EHR Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinGen Allele Registry</td>
<td>Identifiers for genetic variants</td>
<td>Genetic Lab Review/Entry</td>
</tr>
<tr>
<td>ClinGen Curated Summaries</td>
<td>Gene-drug interactions and genetically caused diseases</td>
<td>Problem List Review/Entry</td>
</tr>
<tr>
<td>ClinVar</td>
<td>Clinical significance of known genetic variants</td>
<td>Genetic Lab Review/Entry</td>
</tr>
<tr>
<td>CPIC</td>
<td>Clinical guidelines on genetic testing and treatment related to pharmacogenomics</td>
<td>Medication List Review/Entry</td>
</tr>
<tr>
<td>GeneReviews</td>
<td>Genetic disease knowledge management</td>
<td>Problem List Review/Entry</td>
</tr>
<tr>
<td>Genetics Home Reference</td>
<td>Consumer health information</td>
<td>Genetic Lab Review/Entry</td>
</tr>
<tr>
<td>Genetic Practice Guidelines</td>
<td>Genetic disease guidelines</td>
<td>Problem and Medication List Review/Entry</td>
</tr>
<tr>
<td>Genetic Testing Registry</td>
<td>Genetic tests and laboratories</td>
<td>Problem List Review/Entry</td>
</tr>
<tr>
<td>PharmGKB</td>
<td>Curated pharmacogenomics knowledge base</td>
<td>Medication List Review/Entry</td>
</tr>
</tbody>
</table>

Development of a prototype EHR with gene sequencing panel reports coupled with infobuttons
The EHR prototype is available at http://lite.bmi.utah.edu/ClinGenDemo/Interface_Project/main.html. Four sections were created in the prototype EHR user interface (Figure 1): Patient Profile, Problems, Medications, and Results. The
Patient Profile tab contains patient demographics, allergies, future appointments, and a history of past appointments. These data were populated from the following FHIR resources respectively: Patient, AllergyIntolerance, Appointment, and Encounter. The Problems, Medications, and Results tabs were populated by Condition, MedicationStatement, and Observation FHIR resources respectively. The Problems and Medications tabs are formatted as lists of buttons that, when clicked, present relevant metadata (e.g., onset date, “last updated”, prescription date). The Results tab is also presented as a list of buttons that, when clicked, show other genetic information such as the clinical significance and a single nucleotide polymorphisms (SNP) identifier (if available), as well as a link to the referenced multi-gene sequencing panel report (Figure 2). Each panel report contains infobuttons next to the gene and variant names. Infobuttons were also integrated into the EHR next to the conditions, medications, and genes/variants. A dropdown list on the top right of the EHR prototype contains the patient names. When a name is selected, the information from the corresponding patient’s JSON file is extracted and presented via the EHR interface.

Figure 1. Main screen of the EHR prototype showing the Profile, Problems, Medications, and Results tabs which can be accessed by clicking on the respective tab headings.

Figure 2. An example of genetic laboratory results as shown in the Results tab. A clinician can click on individual results to view the Clinical Significance, SNP, and a link to the full laboratory report from which each result is retrieved. The infobutton “i” icons can be clicked to send HL7-compliant requests to OpenInfobutton to retrieve information on the clinical interpretation of each result.
Infobutton links included a URL-based request to OpenInfobutton including seven parameters, four of which represent the context of the EHR (Table 2). Figures 3 and 4 show examples of the user interface returned by OpenInfobutton after an infobutton is clicked.

**Table 2.** List of parameters included in the HL7-compliant infobutton request sent by the prototype EHR to OpenInfobutton.

<table>
<thead>
<tr>
<th>Infobutton Request URL Segment</th>
<th>Use</th>
<th>Example values</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://service.oib.utah.edu:8080/infobutton-service/infoRequest">http://service.oib.utah.edu:8080/infobutton-service/infoRequest</a></td>
<td>Base URL of the knowledge resource</td>
<td>---</td>
</tr>
<tr>
<td>representedOrganization.id.root</td>
<td>Domain name of requesting organization</td>
<td>ClinicalGenome.org</td>
</tr>
<tr>
<td>taskContext.c.c</td>
<td>Context of EHR task</td>
<td>PROBLISREV, MLREV, GENEREV, VARREV</td>
</tr>
<tr>
<td>mainSearchCriteria.v.c</td>
<td>Search term code</td>
<td>LOINC, or SNOMED-CT codes</td>
</tr>
<tr>
<td>mainSearchCriteria.v.cs</td>
<td>Search term code system</td>
<td>Code system identifier</td>
</tr>
<tr>
<td>mainSearchCriteria.v.dn</td>
<td>Search term</td>
<td>Display name</td>
</tr>
<tr>
<td>informationRecipient.languageCode.c</td>
<td>Language</td>
<td>English, Spanish</td>
</tr>
<tr>
<td>performer</td>
<td>Role of person wanting results</td>
<td>PROV (provider)</td>
</tr>
</tbody>
</table>

**Figure 3.** User interface returned from OpenInfobutton in response to a request for information on the gene “HNF1A” when a provider is reviewing the patient’s lab result list. This example shows a resource that could be used for both clinicians and patients.
Figure 4. User interface returned from OpenInfobutton in response to a request for information on the variant “c.681G>A” when a provider is reviewing the patient’s lab result list. This shows an example of a clinician resource known as “ClinVar.”

Population of the EHR prototype with data from sample laboratory tests and FHIR compliant JSON files
The gene panel prototypes were populated with information about the variant results of the test, the clinical significance, a brief interpretation summary, clinical recommendations, and information about the test itself (Figure 5). References sited within the test are also provided, allowing clinicians to check the sources.

Figure 5. Multi-gene paneling report, which can be accessed from the Results tab. The infobutton links request provider reference information related to genes or variants in the context of genetic test results review.
A Bundle FHIR resource, which holds an embedded list of other FHIR resources, was created for three sample patients, each containing all the resources necessary to populate the EHR prototype (see table 3 for all resources implemented). Table 4 displays the genetic FHIR extensions required to bind each genetic data type, allowing for additional genetic information to be supported outside of the Observation resource. Standard terminologies were used within each FHIR resource as follows: SNOMED-CT for problem list items, medications, and allergies; HGNC for gene names; and LOINC for laboratory tests. The resources for each patient were validated as FHIR compliant by uploading them to the HAPI server, which passes all data through a FHIR-JSON validator before storing to the backend.

Table 3. Links to example JSON FHIR resources.

<table>
<thead>
<tr>
<th>FHIR Resource</th>
<th>Example JSON URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/Patient/cf-149977715321/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/Patient/cf-149977715321/_history/1?_format=json</a></td>
</tr>
<tr>
<td>AllergyIntolerance</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/AllergyIntolerance/cf-1500043653189/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/AllergyIntolerance/cf-1500043653189/_history/1?_format=json</a></td>
</tr>
<tr>
<td>Appointment</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/Appointment/cf-1500043688254/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/Appointment/cf-1500043688254/_history/1?_format=json</a></td>
</tr>
<tr>
<td>Encounter</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/Encounter/cf-1500043817379/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/Encounter/cf-1500043817379/_history/1?_format=json</a></td>
</tr>
<tr>
<td>Condition</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/Condition/cf-1500043739116/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/Condition/cf-1500043739116/_history/1?_format=json</a></td>
</tr>
<tr>
<td>MedicationStatement</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/MedicationStatement/cf-1503330877335/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/MedicationStatement/cf-1503330877335/_history/1?_format=json</a></td>
</tr>
<tr>
<td>Observation</td>
<td><a href="http://fhirtest.uhn.ca/baseDstu3/Observation/cf-1500043993905/_history/1?_format=json">http://fhirtest.uhn.ca/baseDstu3/Observation/cf-1500043993905/_history/1?_format=json</a></td>
</tr>
</tbody>
</table>

Table 4. FHIR Extension references for genetic data types.

<table>
<thead>
<tr>
<th>Genetic Resource Type</th>
<th>Extension URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td><a href="http://hl7.org/fhir/StructureDefinition/observation-geneticsGene">http://hl7.org/fhir/StructureDefinition/observation-geneticsGene</a></td>
</tr>
<tr>
<td>DNA Sequence Variant Name</td>
<td><a href="http://hl7.org/fhir/StructureDefinition/observation-geneticsDNASequenceVariantName">http://hl7.org/fhir/StructureDefinition/observation-geneticsDNASequenceVariantName</a></td>
</tr>
<tr>
<td>Allelic state (Zygosity)</td>
<td><a href="https://www.hl7.org/fhir/extension-observation-geneticsallelicstate.html">https://www.hl7.org/fhir/extension-observation-geneticsallelicstate.html</a></td>
</tr>
<tr>
<td>Amino Acid Change Name</td>
<td><a href="http://hl7.org/fhir/extension-observation-geneticsaminoacidchangename.html">http://hl7.org/fhir/extension-observation-geneticsaminoacidchangename.html</a></td>
</tr>
<tr>
<td>DNA Variant Id (SNP)</td>
<td><a href="http://hl7.org/fhir/StructureDefinition/observation-geneticsDNAVariantId">http://hl7.org/fhir/StructureDefinition/observation-geneticsDNAVariantId</a></td>
</tr>
</tbody>
</table>

Discussion

Clinical Significance
We have developed a prototype EHR with a multi-gene sequencing panel report to test and demonstrate the feasibility of infobuttons to provide clinicians with access to relevant genetic knowledge resources at the point of care. This project demonstrated that several knowledge resources could be integrated with a prototypical EHR through a standards-based approach. It also shows how genetic data and infobutton capabilities can be integrated into an EHR user interface. Health care organizations can follow similar steps to implement this functionality within their EHRS in real health care settings.

We were able to configure eight genetic knowledge resources using LITE, with no need for programming or changes to the underlying OpenInfobutton software. Our experience supports the findings by a usability study conducted by Jing et al., which found that medical librarians were able to configure resource profiles in LITE without any formal training. URL-based, HL7-compliant infobutton requests were also easy to implement into the prototype EHR. Similar experiences have been reported in interviews with implementers of the HL7 Infobutton Standard. Meaningful Use certified EHRS must be able to connect to knowledge resources using the HL7 Infobutton Standard. Therefore, the functionality demonstrated in our prototype EHR can also be implemented within any certified EHR that is configured to use OpenInfobutton. Detailed documentation on how to deploy and use OpenInfobutton is available at http://www.openinfobutton.org/documentation. Three deployment alternatives are available: (1) access a cloud-based
instance, such as the one hosted at the University of Utah (not production-level, available only for development and demonstration purposes); (2) download and install a virtual machine that includes all the OpenInfobutton software; and (3) pull the source code from OpenInfobutton’s GitHub repository and build it within the target deployment environment.

A usability issue to consider when implementing infobuttons in an EHR system is how the infobutton interface is presented to the user. The prototype in this study is coded in such a way to have the interface appear in a new window. However, the infobutton interface can be presented in any way the EHR system developer sees fit, whether in a new window or as a “popup” in the same window. An advantage of showing the interface through a “popup” window is to minimize workflow interruptions. However, showing the interface in a new window allows for multiple infobutton interfaces to be open at once. EHR system implementers will need to consider which option best suits their system when integrating infobuttons.

**Challenges and Limitations**

The vast majority of genetic knowledge resources were not designed for use by clinicians at the point of care and are not yet compliant with the HL7 Infobutton Standard. Therefore, most of the resources required manual mapping of the resources’ API parameters to the Infobutton standard. Those mappings were stored within the resources’ profiles in the infobutton manager and are available for use by any healthcare organization that uses OpenInfobutton.

We faced several interoperability challenges related to the lack of support for standard terminologies in the genetics domain. Knowledge resource APIs and EHR systems do not support gene standard terminologies, such as HGNC.\textsuperscript{19} To address this limitation, we created infobutton requests only with gene names. A similar problem was faced with allele variants. Representing genetic panel reports within FHIR requires the use of a standard terminology to uniquely identify the different kinds of panels. EHR Meaningful Use criteria requires adoption of LOINC to encode laboratory test panels and results. However, LOINC does not currently provide a comprehensive set of concepts for the growing number of available genetic sequencing panels. Therefore, in the *Observation* resources created, a general LOINC code, corresponding with the term “Genetic analysis report”, was used for the terminology code in the *Observation* resources.

Genetic labs still provide their reports primarily in unstructured, PDF format, which imposes a challenge to infobutton integration since discrete genetic results are required for auto-generated infobutton requests\textsuperscript{32}. The sample laboratory reports leveraged to populate the prototype multi-gene sequencing panels were all in PDF format, thus the data had to be loaded into the prototype EHR manually. Similar difficulties will be found in clinical practice until structured, standards-compliant genomic testing reports are more widely implemented.

**Future Work**

Future studies include investigating the usefulness of infobutton responses to clinical genomics inquiries in clinical practice. Ongoing efforts are focusing on continuously expanding, updating and improving OpenInfobutton resource profiles for clinical genomics. Future plans also include deploying access to clinical genomics resources via infobuttons in patient care settings. As mentioned previously, several options are available to facilitate deployment of OpenInfobutton at different healthcare organizations.

**Conclusion**

There is a critical need to provide clinicians with efficient access to clinically actionable genetic information at the point of care. We developed a prototype that demonstrates an approach to help address this need using standards-based functionality that is required for EHR certification in the Meaningful Use program.

**Acknowledgements**

This project was funded by grant number 1U01HG007437-01

Grant name: Clinically Relevant Genetic Variants Resource: A Unified Approach for Identifying Genetic Variants for Clinical Use

**References**


Emergency Department Clinician Perspectives on the Data Availability to Implement Clinical Decision Support Tools for Five Clinical Practice Guidelines

Brian J. Douthit, MSN, RN-BC¹, Rachel L. Richesson, PhD, MPH¹
¹Duke University School of Nursing, Durham, North Carolina, USA

Abstract

Clinical practice guidelines (CPGs) often serve as the knowledge base for clinical decision support (CDS). While CPGs are rigorously created by medical professional societies, the concepts in each guideline may not be sufficient for translation into CDS applications. In addition, clinicians’ perceptions of these concepts may differ greatly, affecting the implementation and impact of CDS within an organization. Five guidelines developed by the American College of Emergency Physicians were systematically explored, generating fifty-one unique clinical concepts. These concepts were presented to two nurses and two physicians, whom were asked to assess and comment on the capture of each clinical concept in the electronic health record (EHR) and the subsequent availability of the data for CDS. Nurses and physicians showed differing perceptions of data availability. These differing perceptions may influence an organizational approach to developing and implementing CDS, potentially informing our understanding of why CDS may not achieve the intended impact.

Introduction

The information seeking behaviors of clinicians perpetuate a need for a growing number of clinical tools⁴⁻⁷. This need is evidenced by the popularity of computerized information repositories (such as UpToDate [http://www.uptodate.com]), the development of InfoButtons, and an emphasis on clinical decision support in both historical and recent contexts (CDS)⁴⁻⁵. While CDS is promoted as an essential strategy for improving the delivery and outcomes of care in the coming era, the development and implementation of CDS tools is still a significant challenge for most organizations⁶⁻⁸. In the face of these challenges, CDS remains as an effective tool to promote care plan adherence, promoting better patient outcomes⁹⁻¹¹.

For CDS to function, an underlying knowledge base is required to influence medical decisions². Oftentimes clinical practice guidelines (CPGs) serve as the knowledge base, thereby instigating the need for their formal representation within a clinical decision support system (CDSS)⁵. CPGs are themselves viewed as a foundational way to standardize care and improve health, but conversion to CDS has been shown to be challenging due to several factors⁴⁻¹³. In particular, the CDSS requires a formal representation of important clinical concepts and the assessments and actions that they drive, as well as a linkage to data elements and patient data as collected in electronic health record (EHR) systems.

This study explores the linkage of patient data to the clinical concepts required for the CDS logic to function as intended, particularly the perceptions of clinician perspectives on data elements within CPGs. This phenomenon is important to understand, as it provides insight into potential barriers affecting development, education, and implementation of CPGs within a CDSS. One portion of data availability seeks to comprehend the cohesion of clinical documentation practices across disciplines, and identify the differences in the perception of available data by the various clinical professions in a formal care setting. Measures for organizational success of CDS often reflect usage and adherence, but the failure of CDS should be more closely considered⁵. The quality of EHR documentation has been shown to vary among practices, which may indicate that accurate and complete documentation could be determined by clinical discipline or individual behaviors⁶. A failure of a CDSS’ implementation may not necessarily result from a poor knowledge base, but potentially a subtle misunderstanding of the availability of the required patient data due to perceptions of the EHR’s ability to capture the clinical process accurately and completely.

The goals of this investigation were to successfully identify clinical concepts within a set of guidelines that could be implemented within a CDSS, examine the similarities and differences of the perception of availability in these concepts among clinicians, and discover phenomenon that could contribute to the understanding of CDS creation, implementation, and maintenance.
Methods

Preparing guidelines to extract clinical concepts for review

The guidelines were selected and processed using a portion of the Shiffman et al.17 methodology for transitioning clinical guidelines into CDSS, with clarification by Tso et al.15 as defined by Table 1. The original guidelines were intended for an automated markup of CPGs, but these steps have been systematically explored to include manual modification15,17. Guidelines were selected and dissected to yield the most granular and operational concepts possible, using the supporting literature of each guideline as a reference when applicable. Each of these concepts were presented to the clinicians, and were asked questions regarding their perceptions of how each concept is represented in the EHR. The responses were categorized and compared to draw preliminary conclusions on the perceptions of data availability among these individuals.

For the Select Guidelines step, the Choosing Wisely (http://www.choosingwisely.org) guidelines were queried, and the first set of guidelines developed by the American College of Emergency Physicians (http://acep.org) were selected15. The following five Choosing Wisely guidelines by the American College of Emergency Physicians (http://www.choosingwisely.org/societies/american-college-of-emergency-physicians/) were selected due to their breadth of clinical concepts and interest to the research team:

1. Avoid computed tomography (CT) scans of the head in emergency department patients with minor head injury who are at low risk based on validated decision rules.
2. Avoid placing indwelling urinary catheters in the emergency department for either urine output monitoring in stable patients who can void, or for patient or staff convenience.
3. Don’t delay engaging available palliative and hospice care services in the emergency department for patients likely to benefit.
4. Avoid antibiotics and wound cultures in emergency department patients with uncomplicated skin and soft tissue abscesses after successful incision and drainage with adequate medical follow up.
5. Avoid instituting intravenous (IV) fluids before doing a trial of oral rehydration therapy in uncomplicated emergency department cases of mild to moderate dehydration in children.

Table 1. Shiffman et al.17 steps as clarified by Tso et al.15

<table>
<thead>
<tr>
<th>Decision Category</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select Guidelines</td>
<td>Choice of specific guidelines and choice of specific recommendations within the selected guidelines to be implemented</td>
</tr>
<tr>
<td>Atomize</td>
<td>The process of extracting and refining single concepts from the narrative text recommendations</td>
</tr>
<tr>
<td>Deabstract</td>
<td>The process of adjusting the level of generality at which a decision variable or action is described to permit operationalization</td>
</tr>
<tr>
<td>Disambiguate</td>
<td>The process of establishing a single semantic interpretation for a recommendation statement</td>
</tr>
</tbody>
</table>

To prepare the five guidelines chosen for this study, the Select Guideline, Atomize, Deabstract, and Disambiguate steps (Table 1) were utilized to create the most granular clinical concepts allowed by the supporting literature of the guidelines to be tested with the clinicians17. The Atomization and Deabstraction steps were repeated as needed to clarify any vague or ambiguous concepts (see Figure 1)17.
Procedures for clinician review of clinical concepts from guidelines

Once the concepts were finalized, three questions were developed to explore the breadth of data availability of each concept in the form of a ‘yes’ or ‘no’. The addition of a comments section was added to provide clarification by the clinician of their choice of answers if needed. The following questions were proposed to the participants of each concept:

1. Is this data captured as discrete data in the EHR?
2. Is this data possibly captured solely as free-text?
3. Is this data found in multiple places in the EHR?

Nurses and physicians were recruited from the Duke University Medical Center (DUMC) emergency department on a volunteer basis. The goal of recruitment was to include a total of four participants; two nurses and two physicians. Inclusion would be met on the basis that the clinician has current experience with the DUMC EHR, and is currently practicing in DUMC emergency department. Four participants were recruited: two nurses and two physicians volunteered and met inclusion. Participants were interviewed individually and asked to review each of the final fifty-one clinical concepts, answering the three questions for each concept. They were given ample time to answer each question with clarification provided when prompted. All questions were answered completely in the form of ‘yes’ or ‘no’. Notes were taken on the context of the answers when provided in the comments section.

To better understand the data availability perceptions of each clinician, analysis of the data consisted of an exploration of discipline specific uniformity, nurse-physician agreement, and nurse-physician disagreement between the two groups. To assess discipline specific uniformity, nurse consensus was first determined by calculating the percentage of consensus for each of the questions. For instance, if both nurses reported that “patient age” is captured as a discrete data in the EHR, that would be considered consensus. Similarly, this was done for the physicians. Nurse-physician agreement was assessed by determining the frequency that all participants (both nurses and both physicians) answered identically for a particular concept. For instance, if both nurses and both physicians reported that “patient age” is not possibly captured solely as free-text in the EHR, that would be considered nurse-physician agreement. Lastly, nurse-physician disagreement was determined by calculating the frequency that both nurses answered ‘yes’ for a concept, while both physicians answered ‘no’, or vice versa.
Analysis of findings from clinical review

The degree of consensus between and across nurses and physicians was determined. Nurse and physician consensus were calculated with corresponding Cohen’s Kappa. Consensus was computed by determining the percentage of concepts that were answered unanimously by both participants in their respective groups. Agreement and disagreement between nurses and physicians were also calculated. Similarly, the percentages of concepts that were answered unanimously by all four participants (two nurses and two physicians) were determined for nurse-physician consensus, while nurse-physician disagreement was calculated as the percentage of concepts that were answered as ‘yes’ by both nurses and ‘no’ by both physicians (or vice versa).

Results

Description concepts in guidelines

Twenty-four concepts were identified and categorized from the five guidelines by the first round of Atomization (“Original Count” in Table 2). The phases of CDS were conceptualized for each concept, identifying “Trigger” criteria, “Inclusion and Exclusion” criteria, and “Recommendations” for information or action (“Original Count” in Table 3). Fifty-six concepts were identified by the remaining iterations of Deabstraction and Disambiguation of the Shiffman et al. process, with five being removed due to identical or nearly identical phrasing (“Final Count” in Table 2). Those in the CDS phase of recommendation were removed from the final product of concepts, as these concepts relate to organization-specific implementation and would not necessarily be impacted by clinician perception (“Final Count” in Table 3).

Table 2. Count of guideline concepts by clinical concept type

<table>
<thead>
<tr>
<th>Concept Type Name</th>
<th>Original Count</th>
<th>Final Count</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1</td>
<td>1</td>
<td>Criteria regarding a patient’s age</td>
<td>Pediatric patients</td>
</tr>
<tr>
<td>Diagnosis/Finding</td>
<td>8</td>
<td>33</td>
<td>Documentation surrounding a clinical finding or formal diagnosis</td>
<td>A history of bleeding disorders, or delayed capillary refill</td>
</tr>
<tr>
<td>Lab</td>
<td>2</td>
<td>1</td>
<td>Regarding a laboratory order</td>
<td>A wound culture is ordered</td>
</tr>
<tr>
<td>Medication</td>
<td>3</td>
<td>3</td>
<td>Regarding a medication</td>
<td>The ordering of an antibiotic</td>
</tr>
<tr>
<td>Meta Process/Evaluation</td>
<td>3</td>
<td>6</td>
<td>Reference to a formal or non-formal clinical process in addition to any evaluation that a clinician may conduct</td>
<td>Determining end-of-life needs for the patient</td>
</tr>
<tr>
<td>Other Order</td>
<td>2</td>
<td>1</td>
<td>An order not pertaining to any other category</td>
<td>A do not resuscitate status</td>
</tr>
<tr>
<td>Procedure</td>
<td>5</td>
<td>6</td>
<td>Regarding diagnostic to therapeutic clinical procedures</td>
<td>Insertion of a Foley catheter</td>
</tr>
</tbody>
</table>

Table 3. CDS phase count

<table>
<thead>
<tr>
<th>CDS Phase Name</th>
<th>Original Count</th>
<th>Final Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigger</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Inclusion and Exclusion</td>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td>Recommendations</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>
**Assessment of clinician consensus**

Discipline specific consensus, nurse-physician agreement rates, disagreement rates, and Cohen’s kappa scores are reported in Table 4. Nurse consensus was the same for questions 1 and 3 at 78%, but varied greatly for question 2 at 57%. Interestingly, physician consensus remained identical for all questions at 78%. Nurse-physician agreement dropped noticeably, at 47%, 35%, and 55% for each question respectively. The nurse-physician disagreement (i.e. both nurses answer ‘yes’ and both physicians answer ‘no’, or vice versa) was 12% and 10% for questions 2 and 3 respectively, with a larger rate of disagreement (18%) for question 1. The Cohen’s kappa statistic varied considerably across the nursing and physician groups, with physicians showing a higher degree of interrater reliability on questions 1 and 3, and nurses showing the highest interrater reliability on question 3. The highest level of agreement (moderate agreement18) was reached by physicians ($\kappa =0.56$) on question 1 and by nurses ($\kappa =0.50$) on question 3.

**Table 4.** Consensus and disagreement of nurse and physician assessments of data availability

<table>
<thead>
<tr>
<th>Question</th>
<th>Nurse Consensus</th>
<th>Nurse Cohen’s Kappa $\kappa$</th>
<th>Physician Consensus</th>
<th>Physician Cohen’s Kappa $\kappa$</th>
<th>Nurse-Physician Agreement</th>
<th>Nurse-Physician Disagreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is this data captured as discrete data in the EHR?</td>
<td>78%</td>
<td>0.24</td>
<td>78%</td>
<td>0.56</td>
<td>47%</td>
<td>18%</td>
</tr>
<tr>
<td>2. Is the data possibly captured solely as free-text?</td>
<td>57%</td>
<td>0.22</td>
<td>78%</td>
<td>0.37</td>
<td>35%</td>
<td>12%</td>
</tr>
<tr>
<td>3. Is this data found in multiple places in the EHR?</td>
<td>78%</td>
<td>0.50</td>
<td>78%</td>
<td>0.12</td>
<td>55%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Each question is graphically displayed (Figure 2, 3, and 4) to compare the nurse-physician agreement and nurse-physician disagreement count of concepts in each category. Multiple trends can be seen in these representations. Of note, the two groups disagreed solely surrounding discrete data (question 1) categorized as a diagnosis/finding (Figure 2). Aside from age, both groups also disagreed on question 2 regarding procedure data at a significant rate (Figure 4). All instances of medication data’s representation in free text (Figure 3) were in consensus by both groups, with a drop in agreement rate (67%) regarding discrete medication data capture (Figure 1).

**Figure 2.** Question 1 (Is this data captured as discrete data in the EHR?) - concept count and percentages by discipline consensus and disagreement

**Note:** Percentages are calculated using the final count of concepts for each category (Table 2)
For CDS to operate as intended, the relevant clinical data must be available to the CDS application. Currently, there is no standard practice for assessing the availability of data from EHR systems, including which users or clinical experts are most able to make this assessment. From this sample, it is suggestive that nurses and physicians may not have the same perceptions of data availability in the EHR. The first question (*Is this data captured as discrete data in the EHR?*) asked of the participants examined their perceptions of discrete data. While natural language processing is beginning to create possibilities for free-text data, discrete data are still seen as more favorable for CDS\(^{19}\). Both physicians and nurses had consensus at the same rate within their groups regarding this question, but cross-discipline consensus was achieved less than half the time. It is also important to note that this question yielded the highest polarization of disagreement, where 18% of the concepts had opposite nurse-physician consensus. Differences in perceptions of discrete data availability is important to understand for CDS implementation, whether it exists between or within groups.
While the consensus of each question (78%, with the exception of question 2 for the nurses) seems similar for both groups, the Cohen’s Kappa highlights areas where perceptions of data availability may not be as comparable if data were collected from a different (or larger) sample. While a firm conclusion cannot be drawn, the perceptions of discrete data availability within the EHR appears to be variable. The addition of questions regarding free-text and multiple data locations helps to better frame this phenomenon with more context from the user’s perspective.

Data availability drives documentation practices. If perceptions differ, data that could have been captured as discrete data may be documented within a free-text note. If different disciplines show a consistent contradictory understanding of what information is available in the EHR, it may be due to issues in EHR understandability, differences in EHR education and exposure, differences in physician and nurse interfaces, or a variety of other human-computer factors. In complex medical documentation systems, it is difficult to ascertain the degree of influence of these factors on data availability perceptions. These questions are important to address as CDS implementations rely on data quality and good documentation practices. With the increasingly present paradigm of patient-centered care, it is imperative that documentation is cohesive among all disciplines, as a break in the continuity of documentation lessens the potential benefit of the EHR in clinical practice. If a phenomenon is understood and documented differently in the EHR, there exists a barrier to truly achieving automated patient-centric CDS. If an organization understands these differences, steps to mitigate these practices could take place by either the consideration of the data sources, or clinician education.

EHR data is captured in various formats, and it would not be realistic to expect all data to be discrete. Certain aspects of care are more appropriately captured as a narrative, but this presents a challenge when attempting to operationalize certain clinical practice guidelines into CDS. While free text can be appropriate, representing clinical concepts in the EHR with discrete data is currently the most pragmatic approach to maximize data availability and interoperability.

The second question (Is the data possibly captured solely as free-text?) serves two functions. First, it explores the possibility that if a concept is not captured as discrete data, it is at least captured in free-text format. Second, if data is able to be captured as discrete data, the clinicians may identify instances where it is not. From the data collected, there are differences in the consensus of this question between groups. The implications of knowing these responses from the perspective of CDS implementation could help guide data sources and education effort. The second function of this question may also serve to better understand the failure of a CDS implementation. Inappropriate use of free-text space, which is more likely in fast-paced environments such as the emergency department, may produce a low utilization rate of CDS, hampering patient outcomes.

In addition to the formats of structured and unstructured data, we examined different clinician views of data availability in terms of its readiness for CDS. Documentation practices are not restricted to a binary understanding of discrete data documentation versus free-text data documentation, but instead represent a more complex facet of CDS implementation. The representation of data, even in the preferred format, can complicate the implementation and expected results of CDS if other dynamics of the EHR are not considered.

The third question (Is this data found in multiple places in the EHR?) explores the idea that data could possibly be documented in multiple locations of the EHR. This question had the lowest rate of disagreement between groups at 10%. While the participants were generally in consensus, there appears to be the most discordance in the documentation of procedures (Figure 4). While it is not readily apparent why, additional questions may be asked of the participants to clarify this phenomenon. If the documentation of a procedure is done multiple times, that is less harmful than the procedure being documented once in varying locations dependent on the clinician. Again, this would be beneficial to know from an organization’s perspective as the selection of data sources could be affected by the documentation practices and perceptions of the clinicians.

Limitations of this study exist largely in the sampling methods and sample size. A convenience sample of participants was taken from the DUMC emergency department in Durham, North Carolina. While this was done purposefully to ensure that each participant answered questions regarding a shared experience, the results’ generalizability to other locations and clinical areas is limited. The small convenience sample of two nurses and two physicians also limits the results, thus a larger number of participants would be recommended for replication of this study. The sample of five guidelines was purposeful to limit the number of questions the participants would have to answer, but a larger sample of guidelines would provide more context, especially if trialed in other clinical areas.

Despite these limitations, our methods built on existing methodological processes and demonstrated a feasible approach to answering questions regarding data availability. We wanted to assess differences in clinician perceptions of data availability by way of a brief interview, thus the design of the three questions was purposely crafted to provide maximum insight with minimal burden on the participants. Due to the short duration of the interviews and possibility
for the questions to be answered using an online survey, it would not be unreasonable for an organization to utilize this type of data collection to inform CDS implementation. The generalizability of these questions lessens the need for modifications to be used in other institutions and clinical specialties.

Conclusion

The factors of clinician perception of data availability are underappreciated when considering CDS creation, implementation, and maintenance. The implications of this study could modify organizational approaches to CDS and EHR training and follow-up. While this study is limited by its sample size and number of participants, the methods may be replicated in various settings with different CPGs and a larger sample to gain more dependable results. If expanded upon, it would be recommended to include perceptions from a data analyst, as they often serve as the primary source of expertise for CDS development in an organization. Juxtaposition of an analyst’s perceptions against the clinician’s perceptions would allow for more in-depth discussion of organizational strategy surrounding this phenomenon. While this preliminary data supports the notion that different clinical professions have differing understandings of data availability, the reason is not completely understood. As more examinations on this phenomenon are conducted, a better approach to these problems may be developed, leading to stronger CDS implementations and better patient outcomes.

Acknowledgements

This work was supported by the Robert Wood Johnson Foundation and the National Institutes of Health grant 1R15LM012335-01A1, Quantifying System and Data Readiness for Automated Clinical Decision Support. Thank you to Dr. Guilherme Del Fiol, Dr. Kensaku Kawamoto, Dr. Catherine Staes, and Fredo Louis of the University of Utah; and Dr. Traci Thoureen, Dr. Neel Kapadia, Ann White, and Melissa Robinson of the Duke University Hospital for making this work possible.

References

Trend Analysis of Aggregate Outcomes in Complex Health Survey Data

Wesley M. Durand, Paul C. Stey, PhD, Elizabeth S. Chen, PhD, Indra Neil Sarkar, PhD, MLIS
Alpert Medical School and Center for Biomedical Informatics, Brown University, Providence, RI

Abstract

Public health and clinical practice pattern trends are often analyzed using complex survey data. Use of statistical approaches that do not account for survey design predisposes to error, potentially leading to resource misdirection and inefficiency. This study examined two techniques for analyzing trends in complex survey data: (1) design-corrected logistic regression and (2) jackknife re-weighted linear regression. These approaches were compared to weighted least squares regression, as well as non-design corrected techniques. Data were obtained from NEISS, a complex survey of emergency departments that can be weighted to produce national estimates of injury occurrence. Trends were analyzed in rug-related injuries among male versus female patients ≥65 years of age. All design-corrected techniques performed comparably in assessment of trend within sex-based subgroups. In almost all cases, design-corrected approaches contrasted profoundly with standard statistical techniques. Future analyses may employ these design-corrected approaches to appropriately account for estimate variance in complex survey data.

Introduction

Analysis of complex survey data often requires specialized statistical software or packages. Such procedures utilize survey design parameters to calculate variance associated with estimates from these data (i.e., design-corrected analysis). Use of standard statistical procedures assuming simple random sampling to analyze survey data predisposes to error. Studies of public health trends and clinical practice patterns often utilize complex survey data. Accordingly, analyses utilizing non-design-corrected statistical techniques risk misguided conclusions, potentially leading to resource misdirection and inefficiency.

The National Health and Nutrition Examination Survey (NHANES; CDC), National Inpatient Sample (NIS; HCUP, AHRQ), and National Electronic Injury Surveillance System (NEISS; CPSC, CDC) are complex surveys commonly employed in biomedical and public health research. The longitudinal nature of these datasets is particularly useful for evaluating trends in event occurrence over time. In these analyses, the outcome is often an aggregate within levels of the variable over which trend analysis is desired (e.g., number of liver transplants per year, or aggregate cost of coronary artery bypass graft procedures per year). The requirement to utilize survey design-corrected procedures, however, limits the array of tools available to researchers. This issue is compounded by significant inter-software variability in available procedures and functionality. Accordingly, when examining aggregate outcome trends, many studies in the biomedical literature have either utilized non-design-corrected statistical approaches (e.g., standard linear regression or Cochran Armitage trend tests), refrained from statistical trend analysis, or circumvented direct trend analysis through a variety of methods.

Weighted least-squares regression is a technique for regression analysis of survey data, as published by Gillum, Graves, and Jean in 1996. While widened availability of survey analysis software has reduced use of this technique for continuous, categorical, and count outcomes, select researchers have applied it to trend assessment of aggregate outcomes in recent years. This approach, however, has several disadvantages. It assumes that standard errors are derived from a sufficiently large sample such that they are without sampling error; this may not always be the case. The magnitude of the regression estimate under weighted least squares may differ from that observed without weighting. Finally, the approach may require ad hoc calculation, complicating “out of the box” usage, particularly when analyzing variable interactions.

This study aimed to develop and evaluate approaches for trend assessment of aggregate outcomes in complex survey data that can be easily applied across a broad array of statistical software. Here, the use of two methods is described: (1) design-corrected logistic regression and (2) jackknife re-weighted linear regression. These methods are compared to an existing design-corrected method (weighted least squares regression) as well as non-design-corrected methods.
(standard linear regression and the Cochran Armitage trend test). The impact of using these techniques is illustrated with data from NEISS, analyzing trends in rug-related injuries among patients ≥65 years of age.

**Methods**

**Data Source:** Full-year NEISS data (1999-2015) were abstracted from the Consumer Product Safety Commission query builder ([https://www.cpsc.gov/cgibin/NEISSQuery/home.aspx](https://www.cpsc.gov/cgibin/NEISSQuery/home.aspx)). The NEISS is a stratified sample of ~100 hospital emergency departments (EDs) that can be weighted to produce national estimates of product-related injuries. EDs are divided into strata based on hospital size. Year was treated as an additional stratification variable, such that size-based strata were temporarily separated; this replicates the NEISS’s annual sampling approach.

**Patient Selection:** Records were included based on presence of: a product code for “rugs” (codes 0612, 0676, 0613); patient age ≥65; and diagnosis code for fracture (code = 57).

**Variables:** The primary dependent variable in this analysis was rug-related fracture occurrence among elderly patients. Independent variables were sex and year.

**Statistical Analysis:**

**Design-Corrected: Logistic Regression:** The design-corrected logistic regression approach utilized a scheme for dataset preparation to treat an aggregated outcome as binary. The overall analysis approach is detailed in Figure 1.

![Diagram](https://via.placeholder.com/150)

**Figure 1.** Dataset Creation Strategy for Design-Corrected Logistic Regression.

First, a dataset containing dummy records for each primary sampling unit (PSU) by year was created and merged with isolated patient records, ensuring that all PSUs were represented in the analysis. Patient cases were assigned “cohort=1” and dummy PSU records were assigned “cohort=0” for domain analysis. Note that this step would be required in any analysis seeking to correctly calculate estimate variance, as not all PSUs may have contributed patient records to the analysis.

Subsequently, a denominator dataset was created containing observations for each combination of variables over which trend analysis would occur; in our case, all year*sex variable combinations. These observations were assigned an arbitrary but consistent stratum (stratum = P) and PSU (psu = 999) values. By placing these observations in “lonely PSUs”, estimates from these observations do not add to the overall variance of the sample. Observations were weighted as the difference between an arbitrary but consistent large number (in our case, K = 1,000,000) and the actual observed number of injuries for that year*sex combination. Importantly, patient cases were assigned “event=1” and denominator dataset observations were assigned “event=0”. Denominator dataset observations were also assigned “cohort=1”.

Design-corrected logistic regression was accomplished with the *survey* R package, utilizing “event” as the binary dependent variable and sex (reference = male), year, and an interaction term as independent variables, as illustrated in Equation 1. A domain analysis was performed for observations with “cohort=1”. A custom linear hypothesis test was requested to determine the effect of year among females.
Equation 1: \( \logit(sex, year) = \beta_{\text{intercept}} + \beta_1 \times sex + \beta_2 \times year + \beta_3 \times sex \times year \)

where annual change among males is represented as \( \beta_{\text{year, males}} = \beta_2 \), the difference in annual change among females versus males as \( \beta_{\text{year, \Delta}} = \beta_3 \), and the annual change among females as the linear hypothesis test of \( \beta_{\text{year, females}} = \beta_2 + \beta_3 \)

Accordingly, the modeled absolute number of events can be expressed as a function of the probability of event occurrence, calculated according to Equation 2:

Equation 2: \( N(sex, year) = p(sex, year) \times K = \frac{\logit(sex, year) - K}{1 + \logit(sex, year)} \times K \)

where \( \logit(sex, year) \) is calculated according to Equation 1, and K is the large number used in weighting denominator dataset observations.

Design-Corrected: Jackknife Re-weighted Linear Regression: In contrast to the design-corrected logistic regression approach, the jackknife re-weighted linear regression method involved primarily computational steps. The unique hospital dataset and rug-related injury datasets were combined for jackknife re-weighted analysis (the denominator weight dataset was excluded).

The jackknife approach to complex survey analysis is available as an alternative to Taylor Series linearization in many statistical packages, though, to our knowledge, there is no literature illustrating use in analyzing aggregate outcomes. In this approach, the analysis is repeated for R replicates, where R is equal to the number of PSUs in the sample. For each replicate \( r \), a sequential PSU is selected for deletion from the sample, and observations from the remaining PSUs within the donor stratum (i.e., the stratum of the deleted PSU) are re-weighted according to Equation 3a. A standard linear regression for number of events by year, sex, and their interaction was conducted for each replicate, defined in Equation 4. Estimate magnitude was calculated as the arithmetic mean of replicate estimates, and variance was computed according to Equation 3b. All calculations for the jackknife re-weighted linear regression approach were conducted in Julia 0.6.0, using the RCall.jl package for regression and linear hypothesis testing.

Equation 3a: \( w_{ij}^{(r)} = \frac{w_{ij}}{n_{hr} - 1} \) \( n_{hr} = \frac{1}{\alpha_r} \)

where \( w_{ij} \) is the weight for the \( j \)th member of the \( i \)th PSU, and \( n_{hr} \) is the number of PSUs in the donor stratum \( h \) for replicate \( r = 1, 2, 3... R \).

Equation 3b: \( V(\beta) = \sum_{r=1}^{R} \alpha_r (\beta_r - \beta)^2 \)

where \( \beta \) are the estimated regression coefficients from the full sample for \( \beta \), and \( \beta_r \) is the regression coefficient from the \( r \)th replicate

Equation 4: \( N(sex, year) = \beta_{\text{intercept}} + \beta_1 \times sex + \beta_2 \times year + \beta_3 \times sex \times year \)

where annual change among males is represented as \( \beta_{\text{year, males}} = \beta_2 \), the difference in annual change among females versus males as \( \beta_{\text{year, \Delta}} = \beta_3 \), and the annual change among females as the linear hypothesis test of \( \beta_{\text{year, females}} = \beta_2 + \beta_3 \)

Design-Corrected: Weighted Least Squares Regression: Estimates and standard errors were computed using the survey package for all year*sex combinations. As described by Gillum, Graves, and Jean, regression coefficients and standard errors were calculated according to Equations 5a and 5b. Weighted least squares regression was implemented in Julia 0.6.0.
Equation 5a: \[ \beta = \frac{\sum w_i x_i y_i - (\sum w_i) (\bar{x})(\bar{y})}{\sum w_i x_i^2 - (\sum w_i) \bar{x}^2} \]

Equation 5b: \[ S_\beta = \sqrt{\frac{1}{\sum w_i x_i^2 - (\sum w_i) \bar{x}^2}} \]

where, \( w_i = 1/S_i^2 \), \( \bar{x} = \sum w_i x_i / \sum w_i \), \( \bar{y} = \sum w_i y_i / \sum w_i \)

For inter-group analysis of trend, the difference between the two groups (i.e., number of injuries among females minus number of injuries among males) was modeled as a function of year. Regression estimate and standard error for this value were calculated as in Equation 5a and 5b, with pooled standard deviation determined according to Equation 6:

Equation 6: \[ S_{\text{pooled}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \]

where, \( s_i \) are the standard deviations to be pooled, and \( n_i \) are the sample sizes

**Standard Analysis: Linear Regression:** For comparison, standard linear regression was performed (regression equation identical to Equation 4) for number of events as the dependent variable, with sex, year, and their interaction term as independent variables. Linear regression was implemented in Julia 0.6.0, with use of the RCall.jl package for linear hypothesis testing.

**Standard Analysis: Cochran Armitage Trend Test:** The Cochran Armitage trend test was implemented, comparing linear trend among male versus female injuries without correction for survey design. This approach was implemented in Julia 0.6.0 using the RCall.jl package and the DescTools R package.

**Software:** Dataset cleaning and preparation were accomplished using SAS 9.4 (SAS Institute, Cary, NC) and Julia 0.6.0. Analysis was accomplished as detailed above. Plots of predicted versus observed injury counts were produced in Julia with the PlotlyJS.jl package.

**Results**

**Descriptive Statistics:** In total, 8,005 unweighted records were analyzed, corresponding to a national estimate of 349,605 patients (SE 10,676) (all future statistics are derived from weighted estimates). The cohort was 82.2% female (SE 0.50%) and 17.8% male (SE 0.50%), corresponding to 287,244 (SE 8,923) and 62,361 (SE 2,603) discharges, respectively.

**Design-Corrected Logistic Regression:** The design-corrected logistic regression approach indicated a robust increasing trend in the annual number of rug-related injuries among both males (\( \beta_{\text{year,males}} = 5.120e^{-02}, \text{SE} 9.153e^{-03} \)) and females (\( \beta_{\text{year,females}} = 2.035e^{-02}, \text{SE} 6.901e^{-03} \)). The difference between the two was similarly marked (\( \beta_{\text{year,Δ}} = -3.085e^{-02}, \text{SE} 7.150e^{-03} \)). A plot of modeled annual injury count was slightly non-linear. (Table 1) (Figure 2) (Figure 3)

**Jackknife Re-weighted Linear Regression:** The jackknife re-weighted linear regression approach also exhibited robust positive trends among both males (\( \beta_{\text{year,males}} = 184.8, \text{SE} 33.06 \)) and females (\( \beta_{\text{year,females}} = 337.2, \text{SE} 115.8 \)). Notably, the t-values for these coefficients were very similar to those observed with design-corrected logistic regression. The difference between the two trends, however, was less significant (\( \beta_{\text{year,Δ}} = 152.5, \text{SE} 97.09 \)), and was not similar to that produced above. (Table 1) (Figure 2) (Figure 3)

**Weighted Least Squares Regression:** This approach produced standard errors that were relatively similar to those produced in jackknife re-weighted linear regression, though with coefficients differing to a greater degree (\( \beta_{\text{year,males}} = 153.1, \text{SE} 30.37 \)) (\( \beta_{\text{year,females}} = 219.3, \text{SE} 101.2 \)). The difference observed between the two trends exhibited a comparable standard error to that produced using jackknife re-weighted linear regression (\( \beta_{\text{year,Δ}} = 63.08, \text{SE} 92.80 \)). (Table 1) (Figure 2)
**Standard Linear Regression**: Use of standard linear regression produced coefficients with identical magnitude but differing standard errors as compared to those from jackknife re-weighted linear regression ($\beta_{\text{year, males}} = 184.8$, SE 68.05) ($\beta_{\text{year, females}} = 337.2$, SE 68.05). The results for the difference between the trends was near identical ($\beta_{\text{year, } \Delta} = 152.5$, SE 96.24). (Table 1) (Figure 2)

**Cochran Armitage Trend Test**: The Cochran Armitage test indicated a highly robust difference in trend between the two sexes ($Z = 32.92$). (Table 1)

**Table 1.** Comparison of Select Coefficients for Trend Analysis.

<table>
<thead>
<tr>
<th>Design-Corrected Analysis</th>
<th>Key</th>
<th>$\beta_{\text{year, males}}$</th>
<th>$\beta_{\text{year, females}}$</th>
<th>$\beta_{\text{year, } \Delta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Logistic Regression</strong></td>
<td>Estimate</td>
<td>5.12E-02</td>
<td>2.04E-02</td>
<td>-3.09E-02</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>9.15E-03</td>
<td>6.90E-03</td>
<td>7.15E-03</td>
</tr>
<tr>
<td></td>
<td>$t$ Value</td>
<td>5.594</td>
<td>2.948</td>
<td>-4.315</td>
</tr>
<tr>
<td><strong>Jackknife Re-weighted Linear Regression</strong></td>
<td>Estimate</td>
<td>184.8</td>
<td>337.2</td>
<td>152.5</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>33.06</td>
<td>115.8</td>
<td>97.09</td>
</tr>
<tr>
<td></td>
<td>$t$ Value</td>
<td>5.59</td>
<td>2.914</td>
<td>1.571</td>
</tr>
<tr>
<td><strong>Weighted Least Squares Regression</strong></td>
<td>Estimate</td>
<td>153.1</td>
<td>219.3</td>
<td>63.07</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>30.37</td>
<td>101.2</td>
<td>92.8</td>
</tr>
<tr>
<td></td>
<td>$t$ Value</td>
<td>5.041</td>
<td>2.168</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>Linear Regression</strong></td>
<td>Estimate</td>
<td>184.8</td>
<td>337.3</td>
<td>152.5</td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>68.05</td>
<td>68.05</td>
<td>96.24</td>
</tr>
<tr>
<td></td>
<td>$t$ Value</td>
<td>2.716</td>
<td>4.957</td>
<td>1.585</td>
</tr>
<tr>
<td><strong>Cochran Armitage Trend Test</strong></td>
<td>Z Value</td>
<td>-</td>
<td>-</td>
<td>32.92</td>
</tr>
</tbody>
</table>
Figure 2. Comparison of $t$ Values from Regression Trend Analyses. DC: Design-Corrected.

Figure 3. Observed and Predicted Values from Design-Corrected Logistic and Linear Regression. DCLR, Design-Corrected Logistic Regression.
Discussion

Drivers of U.S. healthcare costs will experience increased scrutiny as health-related expenditures continue to grow. Trend analysis of public health and clinical practice patterns will play a significant role in characterizing these dynamics, and these analyses will often be conducted using complex survey data. Use of statistical techniques that account for estimate variance is critical, particularly when resource-allocation decisions may be made based on the presence or absence of a significant trend.

This investigation illustrated two methods for assessing trends in aggregate outcomes from complex survey data: design-corrected logistic and jackknife re-weighted linear regression. While this study utilized the NEISS dataset, these results are readily generalizable to analysis of other complex survey data (e.g., NIS and NHANES). These approaches were compared to weighted least squares regression—a previously-employed technique for complex survey trend analysis—and standard, non-design-corrected analytic approaches. Both design-corrected and non-design-corrected techniques incorporated survey weights and thus produced identical point estimates for annual event occurrence. Non-design-corrected techniques, however, did not account for survey design parameters and therefore produced skewed values for the variability of these estimates. Ultimately, there was generally greater similarity between design-corrected approaches than when compared to techniques that did not account for survey design.

Comparison of Trends Among Subgroups:

Comparison between logistic and linear regression techniques is more apt with t-values than standard errors. Considering this, all design-corrected techniques produced comparable results for trends among both male and female subgroups, and the logistic and jackknife re-weighted linear techniques yielded profoundly similar values. The design-corrected approaches were highly distinct computationally; such mutual corroboration increases confidence in their validity. All design-corrected techniques exhibited markedly greater t-values for subgroup trends than did standard linear regression. Accordingly, public health and biomedical research that utilizes standard linear regression to analyze trends from complex survey data may be predisposed to misguided conclusions; researchers should instead utilize only design-corrected approaches.

Among techniques employing linear regression, jackknife re-weighted and standard linear regression yielded near-identical point estimates for regression parameters. Weighted least squares regression exhibited generally lower values. This likely occurred due to weighting, with greater importance given to observations with lower standard errors. Higher y-values may therefore be disproportionately penalized, yielding attenuated trend magnitude. This finding has implications for researchers interested in preserving equal-weighting in trend analysis, particularly when their data exhibits a wide range of outcome values.

Comparison of Inter-Group Trends Between Subgroups:

The difference in trend among males versus females varied greatly between design-corrected approaches. For the design-corrected logistic regression approach, this is attributable to the modeling of relative versus absolute annual change (i.e., log-odds of event occurrence versus number of events). As expected, the trend modeled with logistic regression is non-linear. Accordingly, while linear regression coefficients indicated that the annual absolute change in event number among females is greater than among males, logistic regression coefficients showed that the annual relative change was greater among males than among females. Researchers conducting trend analysis should therefore consider whether their hypothesis is best assessed in relative or absolute terms and choose their technique accordingly.

While standard errors were very similar between jackknife re-weighted and weighted least squares regression, regression parameter magnitude was markedly different. As above, this is likely attributable to disproportionate penalizing of larger observations under the weighted least squares approach. Indeed, the difference between jackknife re-weighted and weighted least squares regression coefficients for $\beta_{year,Δ}$ is approximately equal to the difference in differences for the two approaches between $\beta_{year,males}$ and $\beta_{year,males}$.

Notably, jackknife re-weighted linear regression produced comparable results for the difference in trend between males and females as compared to those from standard linear regression. It is possible that this observation is related to features of our dataset, rather than of the method employed; we will test this hypothesis in planned simulation studies. In contrast, all regression-based approaches (including standard linear regression) yielded profoundly different outcomes as compared to a non-design-corrected Cochran Armitage trend test, which showed a highly significant difference between trends among males versus females. This finding has significant implications for future trend analyses of complex survey data. Use of non-design-corrected techniques may not only predispose to error, but the direction of such error (i.e., overestimation versus underestimation) may be dependent on the technique employed.
**Holistic Method Comparison:**

Performance aside, it is important to consider ease of implementation for design-corrected techniques. We observed design-corrected logistic regression to be the most straightforward of those tested, followed by weighted least squares and jackknife re-weighted linear regression. Design-corrected logistic regression required only a single incremental data cleaning step, leveraging existing survey data analysis techniques for further computation. Implementation was straightforward in R, SAS, and Julia, and required very little *de novo* programming. In contrast, the other approaches required a fair degree of customization. Jackknife re-weighted linear regression required not only replicate creation, but also custom variance, standard error, and t-value calculations. Additionally, the considerable number of replicates required for jackknife re-weighted linear regression (1717 replicates in total) made this approach more computationally intensive. Weighted least squares regression required custom regression parameter and variance calculations, as well as pooling of standard errors, techniques with which many researchers may not be familiar.

Nevertheless, design-corrected logistic regression had a notable disadvantage: it modeled the linear annual change in log-odds and therefore assessed non-linear trends in annual event occurrence. As we observed, there may be significant differences between trends modeled in absolute versus relative annual change. While multivariate models are commonplace in epidemiology, researchers desirous of modeling strictly linear absolute annual change may opt for jackknife re-weighted regression. The choice of technique may ultimately be influenced by *a priori* assumptions regarding relationships between dependent and independent variables, as well as patterns observed in exploratory analyses.

**Future Directions**

Further characterization of these techniques will involve simulated survey datasets to assess reproducibility of trend assessment with replicate sampling. Reproducibility will be measured as the distribution of trend t-values as assessed by each technique across replicate samples. If one believes that a method accurately accounts for sampling approach (i.e., is “design-corrected”), then one should believe that the method will produce relatively more consistent results with replicate sampling than would a method that is not design-corrected. We hypothesize that design-corrected techniques will offer superior reproducibility over non-design-corrected techniques.

**Conclusion**

This investigation showed that, with proper dataset cleaning, design-corrected logistic regression is an accurate and easily implemented method to account for variance in trend analysis of aggregate outcomes from complex survey data. In the context of rug-related fractures among the elderly from the NEISS dataset, the results produced with design-corrected logistic regression were highly similar to those from jackknife re-weighted linear regression; the latter may also be employed when researchers are desirous of modeling strictly linear change in absolute quantities. Generally, all design-corrected techniques produced markedly different results as compared to those from standard linear regression and Cochran Armitage trend test. Thus, the results from non-design-corrected analysis may not only lead to misguided conclusions regarding trend robustness, but also manifest in either overestimation or underestimation, depending on the non-design-corrected technique employed. Future analyses of trends in aggregate outcomes from complex survey data should be conducted using only statistical techniques that account for estimate variance. Use of standard statistical procedures predisposes to error and should be avoided.

**Acknowledgments**

This work was funded in part by the Scholarly Concentration Program in The Warren Alpert Medical School of Brown University and National Institutes of Health grant U54GM115677. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

**References**

Predicting Neonatal Encephalopathy From Maternal Data in Electronic Medical Records

Thomas Li¹, Cheng Gao, PhD², Chao Yan, MS², Sarah Osmundson, MD², Bradley A. Malin, PhD², You Chen, PhD²
¹Duke University, Durham, NC; ²Vanderbilt University, Nashville, TN

Abstract

Neonatal encephalopathy (NE) is a leading cause of neonatal mortality and lifetime neurological disability. The earlier the risk of NE can be assessed, the more effective interventions can be in preventing adverse outcomes. Existing studies that focus on intrapartum risk factors do not provide the early prognostic forecasting necessary to prepare healthcare professionals to intervene early in a high-risk NE case. This work used maternal data in a supervised machine learning framework to predict NE events. Specifically, we 1) collected the electronic medical records (EMRs) for 104 NE newborns and 31,054 non-NE newborns and their mothers, 2) trained and tested a regularized logistic regression on imbalanced and high-dimensional EMR data, and 3) discerned important features that could be possible risk factors. The learned model offers prenatal predictions of NE cases with an average area under the receiving operator characteristic curve (AUC) of 87% and identified the most important predictors.

1 Introduction

Neonatal encephalopathy (NE) is a leading cause of infant mortality and long-term neurological morbidity.¹,² The American College of Obstetricians and Gynecologists (ACOG) defines it as a syndrome of disturbed neurological function in the earliest days of life in an infant born at or beyond 35 weeks of gestation, manifested by a subnormal level of consciousness or seizures, and often accompanied by difficulty with initiating and maintaining respiration and depression of tone and reflexes.¹ Historically, the cause of NE was assumed to be asphyxial in nature, and the term hypoxic-ischemic encephalopathy was widely used until the broader term neonatal encephalopathy was recently standardized by the ACOG. The change in terminology demonstrates a paradigm shift and recognition that the syndrome has multiple causal pathways. It was found that 70% of NE cases were most likely attributed to antepartum risk factors like abnormal fetal growth, maternal infection, and socioeconomic status.³,⁴ Recognizing distal risk factors such as these are important because they may exert early pathological damage and, as an artifact, predispose the fetus to NE. If this is the case, then early, preventative interventions can minimize the effects of distal risk factors and decrease the risk of NE.¹

Traditionally, approaches based on the knowledge of clinical experts have been relied upon to learn risk factors associated with NE. Domain expertise is leveraged to determine a set of candidate variables, after which statistical models are applied to ascertain which variables demonstrated an association with NE through a univariate analysis.³ Another study considered over 50 antepartum and intrapartum features and constructed a high performance logistic regression model using eight features that were identified as significantly associated with NE via univariate chi-squared tests.¹⁰ There are, however, several limitations to these previous approaches. First, this type of investigation restricts the study to a small number of variables, making it difficult to identify novel risk factors. Second, since these approaches rely on the expert knowledge, the variables investigated may be subject to selection bias, such that potential predictor variables could be neglected. Third, previous studies focused on assessing the risk of NE only after birth, using variables that often are inaccessible before delivery, such as birth weight and Apgar scores.

To the best of our knowledge, there has not been an investigation into the extent to which NE can be predicted ahead of birth. If preterm predictions could be made, clinicians could be afforded an opportunity to improve the quality of NE care or prevent it from developing. For instance, it has been shown that planned or elective Cesarean delivery is associated with a reduced incidence of infants suffering from NE.¹²⁻¹⁴ To this end, we aimed to leverage antenatal maternal variables to model the risk of NE before birth. Maternal features have been extensively studied.⁵,⁷,¹⁵,¹⁶ However, maternal features were often treated as discrete variables and mixed in with intrapartum and neonatal features to do NE risk assessment. In the context of NE, there has not been a targeted effort to diversely investigate the
Our investigation was motivated by the need for a data-driven approach that models the risk of NE from a broad set of maternal variables. Our goal was to study the predictor variables that can present early in the NE causal pathway while mitigating variable selection bias. To this end, we present a retrospective study of 31,158 mother-infant pairs from over a decade of maternal electronic medical record (EMR) data at Vanderbilt University Medical Center (VUMC)17. The maternal variables were defined by billing codes that represent symptoms, diagnosis, and procedures, as well as demographic information. Given the low incidence of NE, we adopt a bootstrapping strategy to handle the imbalanced data (104 NE cases vs. 31,054 NE controls). We introduce a regularized logistic regression model that ranks the importance of maternal variables in terms of their ability to distinguish NE cases from control.

2 Materials and Methods

2.1 Materials

2.1.1 Dataset

The dataset was drawn from the EMR system at VUMC, which covers over 2.5 million patient records over a 29-year period. All data were de-identified and the study was approved by the Vanderbilt IRB. Table 1 summarizes various features about the population and the balance of the dataset. We identified 31,158 links between mothers and infants. The EMRs of the infants were inspected for the purpose of case identification of NE1. Maternal EMR information prior to delivery was collected to train and test the classification models, as well as the subsequent analysis of risk factors. Specifically, for each delivery event, we collected age, race, ICD-9 and CPT codes assigned before the time of the linked delivery. As shown in Table 1, the average age of mothers at delivery was 23 years old, with the youngest at age 12 and the oldest at age 50. The mothers were 66% White, 17.2% Black and 16.8% some other races. There were 7,860 unique ICD-9 codes and 6,095 unique CPT procedural codes observed in the mothers' EMRs prior to delivery.

2.1.2 Identification of Cases and Controls

In this study, a case corresponds to a mother who had at least one infant who was diagnosed with NE. We identified cases through a two-step process. First, we automatically identified NE candidates from the clinical notes of the delivered infants of the mothers by matching keywords in the natural language. Specifically, we conducted a free text search using regular expressions built on a set of keywords: {HIE, hypoxic-ischemic encephalopathy, hypoxemia, neonatal encephalopathy, asphyxia}. In doing so, we extracted 141 NE candidates. Since matching a keyword does not guarantee the actual incidence of the disease, in the second step, we manually reviewed each of the clinical notes of NE candidates, and with the confirmation of a clinically-knowledgeable expert, we labeled 104 of the candidates as NE and their mothers as cases. One of the primary reasons that the 37 candidates were excluded is that their clinical notes had some expression of a negation of the matched term (e.g., the infant did not suffer from HIE).

We defined a control as a mother who did not deliver any infants diagnosed with NE during a birthing event. As such, controls include mothers whose infants did not match on any term in the keyword list, as well as the mothers of infants who were considered as NE candidates in the first step of case identification but filtered by manual review at the second step. In the end, the number of controls was 31,054. Note that the case identification is based on each delivery event, such that a mother who gave birth to both a NE infant and a non-NE infant(s) and could be labeled as both a case and control (though this would constitute two separate instances). However, we did not observe any mothers that would be labeled as both a case and control.

1The VUMC creates an EMR for each delivered infant.
2.2 Methods

We designed a data-driven framework that consists of 5 core modules: 1) feature construction to characterize each case/control, 2) cohort establishment to handle the imbalanced data, 3) predictive model training, 4) performance evaluation to assess predictive capability, and 5) feature analysis to determine which features were most associated with NE. The 5 modules are depicted in Figure 1. Details for each module are reported in the following subsections.

![Figure 1: A framework for NE prediction and risk factor analysis based on EMR data.](image)

2.2.1 Feature Construction

A feature space from the records of mothers was constructed. For each mother in the cases/controls, we extracted age, race, and the ICD-9 and CPT codes that were documented in the EMR within one year prior to delivery. The billing codes were limited to one year prior to delivery in an attempt to control for variability of observational window duration between cases and controls. This decision is explained in more depth in the discussion section. To reduce the dimensionality of the feature space and mitigate redundancy, we generalized ICD-9 codes by rolling them up the hierarchy as follows. First, all codes that pertain to pregnancy, childbirth, or conditions originating from the perinatal period (e.g., 630.* to 679.*, 760.* to 779.*, and 23.* to 23.9*) were rolled up to one character after their decimal. Second, all other codes were rolled up to the set of characters before the decimal point (e.g., 320.1 and 320.82 both generalized to 320).

As summarized in Table 2, all of the ICD-9 and CPT codes were represented as binary features, indicating the presence or absence of such codes in the EMR. Age and race were represented as ordinal and nominal features respectively. We borrowed the age categories from previous clinical studies on NE\cite{3, 5}. Since only a small number of mothers fell into
race categories other than “black” or “white”, we grouped them into the ”other” category.

2.2.2 Cohort Construction

As mentioned earlier, our dataset is highly imbalanced in terms of cases and controls, which is a concern for several reasons. First, it is widely accepted that learning from an imbalanced dataset leads to models that lack stability and exhibit poor generalization performance\(^\text{18}\). Second, there is a strong possibility that training in this setting will yield a model that simply predicts each newborn as a control (i.e., no NE). The model would achieve both a high accuracy and high false negative rate (i.e., predicting cases as controls) and would, thus, lack clinical utility. To address the class imbalance problem, we adopted an undersampling technique to construct a set of controls that contain the same number of instances as the cases.

Given that a single sample, although constructed randomly, may yield a biased prediction model, we conducted sampling in a bootstrapped manner, such that we composed \(k\) sample sets to evaluate the variance and confidence interval of the learned statistics. Specifically, we sampled 300 sets of 104 controls, each of which was analyzed with respect to the same set of 104 cases. For each cohort, features that appeared in less than 1% of the instances in a single cohort were eliminated from that cohort’s feature space to reduce sparsity and streamline data transfer.

2.2.3 Predictive Models

To train the models and evaluate their performance on our 300 cohorts, we used 75% of the data to perform a stratified sampling within each cohort for 10-fold cross-validation. The remaining 25% of the cohort served as a test dataset. This sampling strategy ensures that the models are both trained and evaluated on the rebalanced data.

During the training phase, we trained a regularized logistic regression model with a mixture of the \(\ell_1\) and \(\ell_2\) penalties, also known as the elastic net penalty, that identified important features. This method has been shown to perform better at various classification problems than both the \(\ell_1\) or \(\ell_2\) penalties alone on input spaces where features greatly outnumber the number of samples. Regularization is designed to limit error due to variance, and thus improve the generalizability of a trained model\(^\text{19}\). This is a highly desirable quality for our use case of training many models on many cohorts from a single population because all the models should perform consistently well, and be robust to the noise in the specific cohort they were trained on.

Regularized logistic regression is able to perform automatic feature selection\(^\text{19, 20}\) by ranking features and selecting a subset of heavily weighted features. In doing so, it eliminates or reduces the impact of less important features by setting their coefficients to zero or close to zero. The coefficients corresponding to features, i.e. the parameters of the model, indicate the importance of features. Based on this metric, feature selection is conducted.

After executing the training and validation module, we generate 300 predictive models with different coefficients, which we represent as \(\{f_1, f_2, ..., f_m\}\) in Figure 1.

2.2.4 Performance Evaluation and Feature Ranking

To evaluate the generalization performance of each of the learned predictive models, we applied each model on its corresponding test dataset (i.e., a group of the remaining 25% samples in each cohort). We report the distribution of the area under the receiver operating characteristic curve (AUC) measure. Moreover, we record precision, sensitivity, and specificity for each model and computed the mean and the 95% confidence interval of AUC, precision, sensitivity and specificity values of all 300 models.
As mentioned above, regularization distinguishes important features\(^2\) from among a large feature space in each classifier. These features are either positive or negative indicators for the presence of NE. To a certain degree, the learned coefficients reflect the relative contribution of the corresponding features to the NE prediction. If 0 represents the non-NE class and 1 represents the NE class, then the greater the weight of a certain feature, the more its presence relates to the NE. In contrast, the more negative a weight is, the more its presence is associated with a non-NE outcome. If a feature is eliminated from a model, then its corresponding coefficient will be zero.

To determine the predictive performance of a feature for NE, we computed the proportion that a feature’s weight covers with respect to the sum of all weights in the model. These proportions were then averaged across all models that utilized that feature, or, in other words, all models that retained the feature with a non-zero coefficient. We refer to this value as the feature importance, which is formalized in Equation 1, where the importance of feature \(x_p\) is computed based on the learned \(m \times n\) weights \(\{\omega_{ik}\}\) from \(m\) classifiers and \(n\) features. We also designed another metric, average frequency, to measure the contribution of each feature to NE classification. Average frequency for feature \(x_p\) is calculated as the fraction of models having a non-zero coefficient for feature \(x_p\) out of all of the \(m\) models. This metric is formalized as Equation 2, where function \(I(\gamma)\) equals to 1 if condition \(\gamma\) is satisfied, 0 otherwise.

\[
\text{Importance}(x_p) = 1 \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{n} \frac{\omega_{pi}}{\omega_{ik}} \tag{1}
\]

\[
\text{Average Frequency}(x_p) = 1 \frac{1}{m} \sum_{i=1}^{m} I(\omega_{pi} \neq 0). \tag{2}
\]

After computing the importance and the average frequency of each feature, we were able to rank features accordingly. Features above a certain rank are regarded as potentially important features, which, may serve as early evidence for either the presence or absence of NE.

3 Results

An average of 796.6 blank sparse features (occurring in less than 1% of examples within a cohort) were eliminated from each cohort. Table 3 shows that after feature and cohort construction steps, cohorts had 213.8 ICD-9 codes on average and 336.8 CPT codes on average, constituting 552.6 features on average including Age and Race. Models had an average of 45.06 features with non-zero coefficients. In other words, an average of 507.54 features were eliminated during model training. In the subsequent sections, we report the results of 1) the balanced performance of 300 predictive models on the NE prediction and 2) the importance of the features.

3.1 The balanced AUC, precision, sensitivity and specificity

For each of the 300 predicted models, we recorded their AUC score, precision, sensitivity, and specificity. The distribution of the predictive models as a function of AUC is shown in Table 5. It can be seen that 89% of the models achieved an AUC larger than 0.8. The mean of the AUC scores is 0.869 and its distribution is shown in Figure 2.

A Kolmogorov-Smirnov Test\(^2\) failed to reject a null hypothesis that the sample distribution of AUC scores satisfies a normal distribution at the 0.05 significance level. The 95% confidence interval for the mean of AUC scores was 0.8625 to 0.8745. The small confidence interval suggests that proposed framework can support NE prediction models with both stable and high predictive performance.

---

\(^2\)The threshold \(|w| > 0\) is used to define the importance of features.
Table 4: Performance of the NE models.

<table>
<thead>
<tr>
<th>Mean</th>
<th>AUC</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8685</td>
<td>0.8080</td>
<td>0.8079</td>
<td>0.8104</td>
<td></td>
</tr>
<tr>
<td>95% Confidence Interval</td>
<td>[0.8625, 0.8745]</td>
<td>[0.8000, 0.8159]</td>
<td>[0.7982, 0.8177]</td>
<td>[0.8014, 0.8194]</td>
</tr>
<tr>
<td>K-S Test</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
<td>Normal</td>
</tr>
</tbody>
</table>

Figure 2: Histogram of AUC scores of 300 models. The normal distribution in red curve is generated based on the obtained mean and the standard deviation.

Our models also achieved an average precision of 0.8080 within a 0.8000 to 0.8159 95% confidence interval, an average sensitivity of 0.8079 within 0.7982 to 0.8177, and specificity of 0.8104 within 0.8014 to 0.8194.

3.2 Features Importance

On average, there were a small number of important features (~50) in each predictive model. Generally, over 90% features were eliminated in the models. To orient the audience, we only report on features that, on average, have an importance value greater than or equal to 0.05, and an average frequency value greater than or equal to 0.2. These two measures suggested 15 features, shown in Figure 3, for further investigation. Intuitively, features such as ICD-9 code 660.0 for obstructed labor by malposition of fetus at onset of labor and 660.6 for a failed trail of labor, make sense as positive predictors and are consistent with the finding that acute intrapartum events can lead to NE\(^1\). Note that the features that appear to represent observations after a birth, such as negative predictors ICD-9: V24 and CPT: 99238, are most likely from a previous delivery the mother had. These features that occur after a mother’s previous deliveries are still considered prenatal features with respect to her subsequent deliveries.

Age range and race category were included in the model to control for variability of age and race between case and controls. Importance and frequency of these potential confounding variables, shown in Figure 4, are relatively minimal, indicating that the distribution of age and race is similar between case and controls.

CPT code 82947 was a positive feature that exhibited a much higher average importance than any other feature and appeared in all 300 models. To verify that the cases were not dominated by this feature, we calculated the number of cases having that code and found only 16 cases that had this feature. Furthermore, we investigated if there were any
<table>
<thead>
<tr>
<th>Feature</th>
<th>Average Importance</th>
<th>Average Frequency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD9: 660.6</td>
<td></td>
<td></td>
<td>Failed trial of labor unspecified</td>
</tr>
<tr>
<td>CPT: 76827</td>
<td></td>
<td></td>
<td>Doppler echocardiography, fetal, cardiovascular system, pulsed wave and/or continuous wave with spectral display; complete. Follow up or repeat study</td>
</tr>
<tr>
<td>ICD9: 660</td>
<td></td>
<td></td>
<td>Obstructed Labor</td>
</tr>
<tr>
<td>ICD9: V72</td>
<td></td>
<td></td>
<td>Special investigations and examinations</td>
</tr>
<tr>
<td>CPT: 88342</td>
<td></td>
<td></td>
<td>Immunohistochemistry or immunocytochemistry, per specimen; initial single antibody stain procedure</td>
</tr>
<tr>
<td>CPT: 90471</td>
<td></td>
<td></td>
<td>Immunization administration</td>
</tr>
<tr>
<td>CPT: 82947</td>
<td></td>
<td></td>
<td>Glucose; quantitative, blood (except reagent strip)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0.05</th>
<th>0.10</th>
<th>0.15</th>
<th>0.20</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>1</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Average Importance</th>
<th>Average Frequency</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD9: 663.8</td>
<td></td>
<td></td>
<td>Other umbilical cord complications complicating labor and delivery, unspecified as to episode of</td>
</tr>
<tr>
<td>CPT: 99395</td>
<td></td>
<td></td>
<td>Periodic comprehensive preventive medicine re-evaluation and management of an individual including examination,</td>
</tr>
<tr>
<td>CPT: 99203</td>
<td></td>
<td></td>
<td>New Patient Visit</td>
</tr>
<tr>
<td>ICD9: V23.9</td>
<td></td>
<td></td>
<td>Supervision of unspecified high-risk pregnancy</td>
</tr>
<tr>
<td>CPT: 99238</td>
<td></td>
<td></td>
<td>Hospital Discharge Day Management Service: face-to-face evaluation and management service with the patient and his/her attending</td>
</tr>
<tr>
<td>CPT: 88307</td>
<td></td>
<td></td>
<td>Surgical pathology, gross and microscopic examination; Uterus, with or without tubes and ovaries, other than neoplastic/prolapase</td>
</tr>
<tr>
<td>ICD9: V24</td>
<td></td>
<td></td>
<td>Postpartum care and examination</td>
</tr>
<tr>
<td>CPT: 99140</td>
<td></td>
<td></td>
<td>Administration of anesthesia complicated by emergency conditions only</td>
</tr>
</tbody>
</table>

|          | 0    | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.50 | 0.75 | 1     |

**Figure 3:** The 15 features with average frequency $\geq 0.2$ and average importance $\geq 0.05$.

codes related with 82947 which appeared in the model. Given that CPT code billing procedure maintains that codes for related procedures are grouped by prefix, we identified codes with 829 as the first three digits and calculated the number of cases having each of them, as shown in Table 5. If these codes are correlated with each other, in the context
of the NE prediction objective, then all of our models will likely be dominated by the 829 procedure group. To further investigate, we measured the correlation between pairs of codes, and the correlation matrix between the CPT codes observed in the 829 group is visualized in Figure 5. From the figure, it can be seen that, surprisingly, little correlation exists between the codes aside from 82951 and 82952.

### Table 5: The number of positive cases with an assigned code from CPT group 829.

<table>
<thead>
<tr>
<th>CPT code</th>
<th>Cases with Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>82947</td>
<td>16</td>
<td>Glucose; quantitative, blood (except reagent strip)</td>
</tr>
<tr>
<td>82950</td>
<td>69</td>
<td>Glucose; post glucose dose (includes glucose)</td>
</tr>
<tr>
<td>82951</td>
<td>14</td>
<td>Glucose; tolerance test, 3 specimens (includes glucose)</td>
</tr>
<tr>
<td>82952</td>
<td>15</td>
<td>Glucose; tolerance test, each additional beyond 3 specimens</td>
</tr>
<tr>
<td>82962</td>
<td>22</td>
<td>Blood glucose by glucose monitoring devices cleared by the FDA for home use</td>
</tr>
<tr>
<td>82977</td>
<td>4</td>
<td>Gamma Glutamyl Transferase</td>
</tr>
</tbody>
</table>

### Figure 4: Average importance and frequency for Age Range and Race Category.

4 Discussion

This work represents a prediction model of NE built around early maternal variables. Our prediction AUC of 87% on a relatively small number of features about an expecting mother is consistent with the hypothesis that cases of NE can be the result of distal risk factors, rather than acute intrapartum events. Since our model was trained with widely used medical billing codes, the same approach is easily replicable by other HCOs that are interested in modeling their maternal population.

Out of the selected features, a form of glucose test (CPT 82947) appeared as a highly important positive predictor. We offer two possible interpretations for this result. First, glucose tests would accompany a case of pregnancy induced diabetes, which could be a real risk factor for NE. Pregnancy induced diabetes is linked to fetal macrosomia\(^2\), a condition of high birth weight due to increased fetal glucose intake, and high birth weight is a well known risk factor for birth trauma and NE\(^1,3,22\). Secondly, glucose tests could be ordered for maternal patients new to VUMC. These new maternal patients may have been identified as predisposed to a high-risk delivery and redirected to VUMC, the premier medical institution in the area. The fact that all cases fall within a years length of encounter, discussed further below, supports this conclusion.

We acknowledge that this is a pilot study and has several limitations. First, the findings may be biased by the way in which high risk cases present to the VUMC. As the leading academic medical center in the area, VUMC receives patient cases that are often more complex or require more specialized care than other institutions to handle. As an artifact, many of the high-risk patients exhibited relatively short observational windows in the VUMC EMR. Figure 6a and 6b illustrates the length of observation history of mothers in the EMR before delivery windows for control and case patients, respectively. Mothers who had babies with NE had one year or less of medical history, which accentuates the concern that positive cases may have complications before receiving care at VUMC, introducing ascertainment bias into the data. Although we attempted to control for this confounding variable by collecting features within one year of the delivery in the controls, the models may, to some degree, be learning the distribution of diagnosis and procedures a patient receives after they are classified as high risk. Second, the number of NE cases in this study was relatively small and may not capture sufficient information of mothers with NE babies to perform prediction. The data observed was
taken from a single healthcare organization, and thus the patient population may be specific to VUMC. Our approach, however, can be applied with the data from any HCO that uses standardized billing terminologies.

Based on our findings and the limitations of the study, we outline the following improvements and extensions for the next iteration of this study. To begin, additional maternal and prenatal features can be investigated such as maternal lab values, in particular glucose values, and fetal ultrasound results. Medication data was previously included in the model, using the same binary feature construction scheme, but these features had low importance in the resulting model and no improvement in prediction was observed. Perhaps a non binary feature encoding may improve the signal received from medication prescriptions. Finally, we hope to understand the billing practices and clinical workflow behind the top features in order infer their potentially causal relationship with NE.

5 Conclusion

NE is a lethal condition that threatens millions of infants with premature death or lifelong morbidity. To date, the pathology of this complex affliction has eluded thorough understanding. Almost all existing NE studies are unable to perform a prenatal risk assessment of NE, making it challenging to adapt efficient interventions to prevent or treat NE in a timely manner. This paper introduces the first data-driven approach to predicting NE before an infant is born, and provides a mechanism to identify high-risk maternal factors for NE. The evaluation relied on a retrospective study of maternal EMR data at a large academic medical center. We considered a representation of the maternal condition centered around symptoms, diagnosis, and procedures, and trained classification models that consistently predict with high performance. However, this study is still preliminary in its scope and size. Further investigation and clinical interpretation of the reported features is required to understand the model, and replication of the study with other populations is needed to arrive at a better picture of the NE pathology. We believe this investigation opens up new and easily replicable approaches to understanding NE, and even may prelude a prediction system for recommending early NE interventions through the electronic health system.

Acknowledgements

This research was supported, in part, by grants T15LM00740, R00LM011933, and R01LM010207 of the National Library of Medicine, National Institutes of Health.
References

18. Haibo He, Garcia E. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering. 2009; 21(9): 1263-84.
i2b2 implemented over SMART-on-FHIR

Nicolas Paris, MSc.1,2,3, Michael Mendis4, Christel Daniel, MD, Ph.D.1,3, Shawn Murphy, MD, Ph.D.1, Xavier Tannier, Ph.D.3,5 and Pierre Zweigenbaum, Ph.D2
1WIND-DSI, AP-HP, Paris, France; 2LIMSI, CNRS, Université Paris-Saclay, Orsay, France; 3INSERM, UMR_S 1142, LIMICS, Paris, France; 4Partners Healthcare In., Boston, MA, USA; 5Sorbonne Universités, UPMC Univ Paris 06, France

Abstract

Integrating Biology and the Bedside (i2b2) is the de-facto open-source medical tool for cohort discovery. Fast Healthcare Interoperability Resources (FHIR) is a new standard for exchanging health care information electronically. Substitutable Modular third-party Applications (SMART) defines the SMART-on-FHIR specification on how applications shall interface with Electronic Health Records (EHR) through FHIR. Related work made it possible to produce FHIR from an i2b2 instance or made i2b2 able to store FHIR datasets. In this paper, we extend i2b2 to search remotely into one or multiple SMART-on-FHIR Application Programming Interfaces (APIs). This enables the federation of queries, security, terminology mapping, and also bridges the gap between i2b2 and modern big-data technologies.

Introduction

Learning Health Systems aim to maximize the potential of large-scale, harmonized data from variable, quickly-developing digital sources such as Electronic Health Records (EHRs), which are emerging as powerful tools to facilitate discoveries that can improve health. Data heterogeneity is one of the critical problems in analyzing, reusing, sharing or linking datasets. With the development of platforms enabling the linking and federation of phenome, genome and exposome data across sites in US,1,2 Europe3,4 or at international scale5 a key challenge is to define harmonized access to heterogeneous EHR-based data.

i2b2 is the de-facto open-source medical tool for cohort discovery and allows healthcare practitioners to easily subset patient data to address research questions. i2b2 has been described as being used by more than 200 hospitals over the world, and the recent migration of i2b2 to GitHub has facilitated development work. The tool is flexible, supporting its own star schema and ontology model as well as exploiting alternative information models such as PCORnet7 and the OMOP common data model8 without requiring changes to the underlying data. Many initiatives have extended the primary goal of cohort discovery, providing functionality to carry out statistical analysis in place, as well as federated queries over multiple centers, and even genomic analytics.9, 10 Well-known tools that extend the i2b2 functionality in this way include SHRINE, INSITE, tranSMART, and TRINETX9, 11–13.

i2b2 and derived solutions do have room for improvement. For example, in terms of data variety, federation tools such as SHRINE, INSITE, and TRINETX are inconsistent in terms of their terminology mapping processes14. When mapping details are provided2, they are time consuming and software specific15. In terms of freshness of the data, Extract Transform Load processes (ETL) feeding traditional relational databases supported by the i2b2 tools (e.g. PostgreSQL, Oracle, MSSQL) are resource consuming, taking considerable amounts of time, maintenance, and disk-space. Though ETL procedures are still feasible these days, the emergence of high-throughput healthcare data and the "Internet of Things" demands the development of new approaches that allow data to be queried in place (i.e. directly within EHRs) or in optimized, dedicated places such as RDF triple stores or NO-SQL databases. The time delta due to data migrations and transformations poses problems of data veracity because the source data is susceptible to be modified in the interval, and multiple transformations are error prone. In terms of data volume, data such as omics, exposomics, imaging or free text notes are more and more produced at hospital level while they still are challenging to store and therefore to analyse. In order for the data to be analyzed properly and efficiently, specialized and dedicated technologies are required. While there have been several engineering attempts to create i2b2 based data-warehouses solutions that work with technologies other than traditional relational databases,16 the cost to create such interfaces is high. The i2b2 star schema model is highly optimized for fast retrieval of lists of patients matching criteria, but it is not intended for statistical analytics or data exploration17. Although there are some bridges with other common data-models such as OMOP, the extended i2b2 architecture is still limited on RDBMS7. The emergence of new technologies
is faster than i2b2’s ability to exploit them. In terms of software accessibility, for example, physicians spend time switching between applications, writing their login and password credentials again & again. Providing these users with the “one login/multiple application” paradigm would optimize the time spent on the computer and thus improve patient care.

The solution explored in this work is to bring the latest accomplishments of the Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) community to i2b2. In particular, to bring the flexibility, the extensibility, the standardization, and the interoperability efforts of FHIR to i2b2. In the domain of patient care, several large-scale efforts have been underway for over a decade with the goal of specifying both the structure and the semantics of patient clinical information in a manner that enables computable semantic interoperability between diverse systems. Although there is no consensus in the medical informatics community regarding a standard patient information model, FHIR specifications are gaining interest and show promise to mitigate the classic site-specific data mapping problem. FHIR is built on lessons from previous standards, including the Reference Information Model (RIM) which became an ISO standard in 2003, and the Clinical Document Architecture designed to express a single clinical document as a message using HL7 version 3 RIM classes. FHIR specifies a RESTful API to access resources. Several initiatives facilitate the adoption of FHIR, including the Argonaut project and the Clinical Information Modeling Initiative (CIMI). SMART-on-FHIR is an open, standards based technology platform that enables innovators to create apps that seamlessly and securely run across the healthcare system. Using EHRs or data warehouses that support the SMART standard, patients, doctors, and healthcare practitioners can draw on this library of applications to improve clinical care, research, and public health. SMART improves the user experience with regard to login details in the way OAuth does for many websites, and there exist currently more than 50 applications that are able to consume FHIR resources in a consistent way.

FHIR and SMART-on-FHIR appear to be good candidates for overcoming several of i2b2’s architectural weaknesses. Several studies have explored how to bridge i2b2 and FHIR. One approach aims at allowing mobile phones to push FHIR resources into the i2b2 star schema. Other approaches allow an existing i2b2 instance to supply their star schema data through a FHIR-API, allowing SMART-on-FHIR applications to run on top of i2b2. In contrast to this prior work our proposal does the exact opposite, in that we build a general interface that allows i2b2 to consume data from any FHIR endpoint. This approach enables clinical datasets to be queried by exploiting FHIR search, Terminology Mapping and SMART Oauth2 security specifications. The aim is not only to bridge the gap between patient care and research communities, but also to open i2b2 to new and improved data types, as well as security and interoperability management in the context of scalable solutions for cross-border and cross-domain networking of data. The ultimate goal of the architecture presented in this paper is to allow multiple institutions to quickly and effectively engage in massive, international cohort discovery studies.

Methods

To meet our objectives, the existing i2b2 traditional query search module (i2b2crc) code source is extended to meet the SMART-on-FHIR API specifications and the FHIR search specifications. Figure 1 shows the overall architecture and how the three-tier i2b2 application integrates with remote institutions. The figure shows how an i2b2 application gives access to users in a SMART-on-FHIR application modality. In this context, users log-in to any SMART application or EHR system just once, and get access to a set of specialized applications, such as i2b2. The architecture allows queries to be combined over multiple endpoints: zero to one i2b2 star schemas and/or zero to many SMART-on-FHIR APIs.

Figure 2 is a detailed UML sequence diagram. The scenario describes a user who runs a query over an i2b2 instance containing both data in its star schema and data present in multiple remote FHIR-API endpoints. The solid arrows represent the specific new methods introduced in this work, the dotted arrows represent the existing i2b2 way of doing. The user first logs into the system with his or her login credentials, which are verified by the i2b2 project management cell (i2b2pm). The i2b2pm gets the user choose from an i2b2 project list according a set of defined roles. After building and running a multiple panel query across different medical domains, the user gets back a patient cohort set by picking concepts from the i2b2 ontology. Some domains are linked to the FHIR endpoints, others are linked to the local star schema. At first, the i2b2crc conducts a patient-set lookup in its local database. Then, for the FHIR-based concepts, the system loops over the following steps against the SMART-on-FHIR API. The i2b2pm first gets its Oauth2 authentication credentials because it is known as a trusted application within the SMART-Auth layer.
The i2b2crc new FHIR query builder then produces the HTTP query according to the FHIR-search specifications and makes an HTTP call to the FHIR-API. This query is extended with coding synonyms defined in the FHIR ConceptMap resources (terminology server part). The resulting query is then translated by the FHIR layer in the local database query dialect to fetch the results. The result is transformed into a FHIR json bundle only containing the information needed (patient_ids in this case). A parsing step extracts the patient_ids. They are then mapped to a unique i2b2 identifier and pushed into a temporary table that integrates all the results. Once looping is done, the i2b2crc applies a new patients security step to only keep those available for the i2b2 project selected by the user. The patient cohort set is finally returned to the user.

<table>
<thead>
<tr>
<th>HTTP request</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET &lt;FHIR-API&gt;/&lt;Resource&gt; ?_elements=&lt;elements&gt;&amp;code=&lt;codes&gt; &amp;date=gt&lt;date_inf&gt;&amp;date=lt&lt;date_sup&gt; &amp;&lt;custom_filter&gt;</td>
<td>Retrieves chosen &lt;elements&gt; from resources optionally matching a date range or/and a list of &lt;codes&gt; or/and a &lt;custom_filter&gt;</td>
</tr>
<tr>
<td>GET &lt;FHIR-API&gt;/ConceptMap ?target-code=&lt;codes&gt; &amp;target-system:in=&lt;code-system&gt;</td>
<td>Retrieves all codes that are mapped to &lt;codes&gt; &amp; &lt;code-system&gt;</td>
</tr>
</tbody>
</table>

**Table 1:** Index of HTTP request templates

*FHIR-search:* FHIR search specifications\(^{25}\) describe how to communicate with a FHIR-API to get back a set of resources matching the criteria in an HTTP query. The present work exploits only the possibility to fetch one type of resource per query. This is sufficient because the traditional i2b2crc allows combining multiple filter predicates by processing each one separately and then uses a deliberation step based on temporary tables. The new i2b2crc query...
builder replaces the SQL queries acting over the star schema to fetch record identifiers (ID) with HTTP calls to a FHIR-API. The latter is then translated into the backend database system. The HTTP calls enabled in this design are presented in Table 1. The first row is the general template that is used, and is compared here with the SQL syntax (SELECT, FROM, WHERE):

**SELECT**: The `<elements>` pattern lists the resource elements that are returned by the FHIR-API. Depending on the user’s choice, patient ID, encounter ID, instance ID or date are retrieved, to respectively provide i2b2 “same patient”, “same encounter”, “same instance”, or “temporal queries” features. The way the i2b2crc retrieves those information from a given resource is described into the i2b2 FHIR config YAML file (see Figure 3).

**FROM**: The `<Resource>` pattern is supposed to be replaced by any existing FHIR standard resource, or any profiled resource (modification of the standard to meet the local institutions constraints). In order to let the user point to the right FHIR resource, the i2b2 traditional ontology table has been reused and populated with the needed information. Table 2 describes how to store the information into the “c_tablename” column.

**WHERE**: Both patterns `<date_inf>` and `<date_sup>` allow filtering the data based on the date range defined by the user at run time. The `<custom_filter>` allows to combine a predefined pattern, such as data status, or a user-defined filter based on a numerical or enumerated value (when enabled by filling the “c_metadataxml” column). The `<codes>` pattern can optionally specify a list of coding (e.g: from SNOMED, LOINC…) by populating the i2b2 ontology “c_basecode” column.
### Table 2: i2b2 ontology adapted for FHIR

<table>
<thead>
<tr>
<th><strong>ontology table columns</strong></th>
<th><strong>Description</strong></th>
<th><strong>Example</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>c_basecode</td>
<td>FHIR code system / code pipe separated</td>
<td>FHIR:<a href="http://loinc.com">http://loinc.com</a></td>
</tr>
<tr>
<td>c_tablename</td>
<td>Resource / Profile pipe separated</td>
<td>Observation</td>
</tr>
<tr>
<td>c_metadataxml</td>
<td>An xml describing datatype (numeric, free text or enumerated) and measure units</td>
<td>cf: i2b2 documentation</td>
</tr>
<tr>
<td>c_dimcode</td>
<td>an optional additional filter</td>
<td>active=true&amp;status=final</td>
</tr>
</tbody>
</table>

### Figure 3: i2b2-FHIR YAML configuration file sample

```
version: dstu3
Patient:
  patientUriPath: $.resource.id
  patientUriField: id
Observation:
  - patientUriPath: $.resource.subject.reference,
  - encounterUriPath: $.resource.context.reference
  - instanceUriPath: $.resource.id
  - datePath: $.resource.effectiveDateTime
  - patientUriField: subject
  - encounterUriField: context
  - instanceUriField: id
  - dateField: effective
```

### FHIR-mapping: The second row of Table 1 describes the HTTP query template to enable terminology mapping. The i2b2crc cell is allowed to fetch synonym codes from the FHIR-Terminology server. As part of the ConceptMap resource, FHIR links a source code to a target with a set of semantic relation such as “equivalent” or “narrower” that characterize the way they relate to each other. The program fetches each mapping pair and only keeps the “wider” and “equivalent” semantic relations. The i2b2-FHIR code expansion exploits this mechanism to query over distinct code systems.

### Outcome measurement: In order to test the FHIR DSTU3 resources compatibility coverage, the HAPI FHIR test server has been used as an endpoint since it contains useful demo datasets with fictitious patient data. The benchmark comparing traditional i2b2 (version 1.7.9) and FHIR-i2b2 was carried out with the same i2b2 “observation_fact” table containing 140 million records in a PostgreSQL 9.6 instance. The FHIR-i2b2 has been set up by implementing HAPI-FHIR server on top of the observation_fact table into an Apache Tomcat 9 webserver, and has been accessed via the FHIR-i2b2 prototype. A FHIR-i2b2 big-data benchmark has been set up by implementing an HAPI-FHIR server on top of the MIMIC-III physiological data table multiplied by 15, and stored in an Apache Hive table distributed over a 5-computer cluster in the Optimized Row Columnar (ORC) format distributed over HDFS. All software used: i2b2, HAPI-FHIR, PostgreSQL and Apache Hive is open-source licensed.

### Results

**Implementation Status:** The design presented below is implemented at 70%. To date, the new i2b2crc query builder is able to query on both a star schema and one remote FHIR endpoint simultaneously. Logical relations between selection criteria represented as multiple i2b2 webclient panels are also possible. The constitution of a “patient_set” can be constrained by dates, values and measurement units and by one or multiple codes. The code expansion based on FHIR terminology mapping is also implemented. A living demo is deployed and a screen-shot presented in Figure 4. The first panel query searched into HAPI FHIR test server for patients with LOINC glucose codes having value lower than 100ml/dl in a year range from 1979 to 2015 and is mixed with the second panel searching for patients having a diagnosis related to circulatory system within the star schema. The resulting “patient_set” contains about...
eight patients.

**Performance:** The performance of i2b2-FHIR has been benchmarked (Figure 5a) versus a traditional i2b2 instance based on the star schema with the same amount of data and configuration (140 million records). The different patient set size corresponds to different criteria selection. The histogram shows traditional i2b2 is 20 times faster than the i2b2-FHIR version. The difference can be explained by the additional steps involved: the fetched result-set is transformed into a json bundle, sent over the network and then parsed. The performance factor tends to decrease with the number of patients matched. The second benchmark (Figure 5b) experimented connecting to an Apache HIVE table on a big-data platform containing 5 billion actual physiological patient data points. The results show that the time spent is under the minute and compatible with i2b2 promises. Moreover, the bar-plots show that the major bottleneck is the FHIR json generation step, mostly the part to produce the json (networking transfer is actually pretty fast). Such quantities of data have never been described to be handled by i2b2 before, since here we approach traditional RDBMS volume limitations. While the traditional i2b2 outperforms the FHIR based i2b2 on modest datasets, the latter opens new perspectives by enabling connections with specialized and optimized database systems.

**i2b2 feature coverage:** The i2b2 querying feature covers filtering patients facts by code, values, dates within an encounter temporal window or even a free sequence of events. By adding new temporal table mechanisms, the present work allows all of those features. Thus it does not limit the existing set of functionalities. The i2b2-FHIR configuration file Figure 3 contains information about the FHIR-API instance, such as its version, and how the resources are implemented. Depending on the kind of cohort set the user wants to extract, patient ID, encounter ID, instance ID or dates are retrieved from the FHIR-API thanks to a jsonPATH description. This then allows to populate the i2b2crc temporary tables. This is how i2b2 deliberation mechanisms can be populated, and the set built. Moreover the i2b2-FHIR implementation keeps backwards compatibility and does not impose FHIR-API usage for implementers that would not require it.

**Security:** A security layer has been proposed and implemented into the existing i2b2crc. A new i2b2 table allows to define which patient are part of which i2b2 project. This security layer is important because with one endpoint containing all patients records, it allows to create multiple projects with subset allowing multiple views on the dataset. In terms of performance, the table might be vertically partitioned and split by project, in order to get stable performance.
Figure 5: (a) Traditional versus i2b2-FHIR performance comparison (on a 150M postgresQL table). (b) i2b2-FHIR performance (on a 5B Hive table)

as the number of projects increases. This mechanism is both compatible with traditional i2b2 and i2b2-FHIR and has been deployed in production at the AP-HP hospitals where it handles more than 200 distinct projects. The Oauth2 security layer has not yet been implemented. The implementation will be inspired by the recent existing implementation such C3-PRO\textsuperscript{22} or SMART-on-FHIR\textsuperscript{23}.

**Extensibility:** To date, the query builder is compatible with the current FHIR DSTU3 version. In the future, it will maintain compatibility with each FHIR release, and also backward compatibilities. The FHIR version of each endpoint is set up in the configuration file (Figure 3). Also it is possible to exploit most aspects of FHIR extensibility, such local profiles, local specific resources or element extensions (through the “c_dimcode” of the i2b2 ontology table). The FHIR access layer has been tested over the HAPI FHIR test server for all resources at least referring to a patient (68 resources), and has a complete resource coverage. The results suggest that the design is flexible enough to query multiple centers with different FHIR implementations at the same time.

**Interoperability:** The FHIR-ConceptMap expansion has been implemented. A set of test mappings have been produced and populated into HAPI-FHIR as a proof of concept. The HTTP query described in Table 1 (row 2) allows to fetch the equivalent codes. While there is still room for improvement, the results open the way for massive and collaborative concept mapping, with a FHIR compatible terminology server. Interoperability is also derived from the FHIR standard resource definition. However, the ability to derive from them and build Profiled Resources is handled by the i2b2-FHIR YAML configuration flexibility together with the i2b2 ontology table, as they are designed to be adapted.

**Discussion**

FHIR abstraction allows designing a mixed architecture based on living EHRs and big-data repositories leveraging massive and unstructured clinical data. One can choose the best technology depending on the expected usage and local specificity of the data. This flexible design allows implementers to define their own i2b2 ontologies. Finally, an i2b2 federation over FHIR is able to bridge multiple FHIR implementations at the same time. The querying benchmarks showed that performance was not really an issue. There is also room for improvements there. High level performance improvement includes parallelizing the queries over both each hospital endpoint and each i2b2 panel. In order to manage heterogeneity in endpoint performance correctly, solutions such background queries have been proposed by SHRINE and could be reproduced. Low level performance improvement includes using more optimized code at both
i2b2 and FHIR endpoint side and at each step of the process: json building, json compression, json parsing. While the FHIR resource subset feature is already used to reduce the size of the bundles, we also will give feedback to the FHIR community to propose a more compact option. Finally the recent FHIR specifications around the GraphQL API language will enable both query optimizations and extensions in the future. For example, GRAPHQL allows mixing multiple FHIR resources and criteria in a single call, and this would enable to combine multiple i2b2 panel criteria into a single FHIR-API call. Such simplification is neither possible with the current i2b2-FHIR implementation nor with the traditional i2b2crc based on the star schema.

By leveraging access to big-data technologies, this opens a new-area of specific solutions (such as temporal-series, text-mining, distributed, graph databases) to manage the diversity, variety and volume of healthcare data such as Genomics, Imaging, or physiological waveforms monitoring. The abstraction provided by the FHIR layer allows plugging new text specific technologies based on Apache Lucene, such SOLR & Elastic search. This will allow clinicians to mine texts as simply as a modern search engine does. Moreover, the interoperability gain from the FHIR interface makes possible to query multiple centers on real-time data (thanks to FHIR API directly plugged on top of EHR) and the same way (thanks to FHIR Concept Mapping ability). The security was enforced and allows multiple sub projects to access subsets of the whole patient database. This provides a restricted access to the needed data by only authorized users accordingly to patient consenting rules.

Several modules have been implemented, some aspects of the design have only been tested as separate modules. The roadmap provides for the development of multiple SMART-on-FHIR endpoints access, OAuth2 implementation, performance improvements and also the release of i2b2 as part of the SMART-on-FHIR apps. Once satisfied with the results, the system should be available in the next releases of the core i2b2. While all resources containing patient reference were tested, there is a need to propose a general mapping between traditional i2b2 objects (patient, visit, provider, observation) and FHIR specific resources (Organization, HealthcareService, Patient, EpisodeOfCare, Condition, Procedure, Medication, MedicationRequest, Observation, DiagnosticReport, ClinicalImpression...). A general algorithm to translate FHIR terminologies into i2b2 ontology will also be investigated, and implemented as complementary software. Last but not least, as a standard way to represent concept mapping, the FHIR ConceptMap resource is a great candidate to centralize and share collaborative work and tools for this major purpose.

**Conclusion**

By bridging current modern solutions in the field of medical data, this work paves the way for improvements that address current Learning Health Systems challenges. The challenges of data federation, data interoperability, data freshness and data security can benefit from both i2b2 experience and FHIR simplification. The challenges of data volume and data variety of medical datasets are indirectly addressed by the FHIR-API abstraction that makes possible the use of powerful and dedicated technology.

In the end, while a tool that is able to bridge international institutions together is likely to emerge, concept mapping between many institutions remains a challenge to be addressed because the use of different languages, different granularities and different medical concepts. While ontology matching is an old research area, it still presents significant challenges to overcome.

**References**


Characterizing Functional Health Status of Surgical Patients in Clinical Notes

Steven J. Skube, MD\(^1\), Elizabeth A. Lindemann, BS\(^2\), Elliot G. Arsoniadis, MD\(^1,2\), Mari Akre, RN, PhD\(^3\), Elizabeth C. Wick, MD\(^4\), Genevieve B. Melton, MD, PhD\(^1,2\)

\(^1\)University of Minnesota Department of Surgery, Minneapolis, MN; \(^2\)University of Minnesota Institute for Health Informatics, Minneapolis, MN; \(^3\)University of Minnesota Physicians, Minneapolis, MN; \(^4\)University of California San Francisco, San Francisco, CA

Abstract

Functional health status is an important factor not only for determining overall health, but also for measuring risks of adverse events. Our hypothesis is that important functional status data is contained in clinical notes. We found that several categories of phrases related to functional status including diagnoses, activity and care assessments, physical exam, functional scores, assistive equipment, symptoms, and surgical history were important factors. Use of functional health status level terms from our chart review compared to National Surgical Quality Improvement Program determination had varying sensitivities for correct functional status category identification, with 96% for independent patients, 60% for partially dependent patients, and 44% for totally dependent patients. Inter-rater agreement assessing term relevance to functional health status was high at 91% (Kappa = 0.74). Functional status-related terms in clinical notes show potential for use in future methodologies for automated detection of functional health status for quality improvement registries and other clinical assessments.

Introduction

Functional health status is often defined as one’s ability to perform daily activities required to meet basic needs, fulfill usual roles, and maintain their health and well-being\(^1\). It is increasingly recognized that a patient’s functional health status is important for determining overall general health and has been used as a factor to estimate pre-operative risk of complications and adverse events\(^2\)\(^-\)\(^5\). Unfortunately, the measurement and documentation of functional health status is often not standardized particularly for front line clinical practice. Often, physicians and other clinicians use a combination of scoring systems, clinical judgement, and physical exam to determine a patient’s functional health status.

There are several tools created to determine a patient’s functional health status. These include tools like the Karnofsky Performance Scale, an observational method of functional health status determination using a 0-100 point scale\(^6\)\(^-\)\(^7\). Functional health status may also be determined through patient or caregiver-completed assessments such as the Patient-Reported Outcomes Measurement Information System (PROMIS)\(^8\). Alternatively, functional health status may be determined through calculations as a composite score. One example of this is the Frailty Index, which measures patient medical problems compared to an age appropriate list of medical problems\(^9\). Metabolic equivalents (METs) are also often used to measure functional capacity, a surrogate of functional health status\(^10\). There are also frameworks and guidelines created to standardize the determination of a patient’s functional health status, such as the World Health Organization’s International Classification of Functioning, Disability and Health\(^11\).

The American College of Surgeons National Surgical Quality Improvement Program (NSQIP) is a quality improvement registry which collects patient data to track post-operative outcomes and complications. This program offers high quality, risk adjusted data that is nationally validated. Data is collected manually by the surgical clinical reviewer (SCR), a trained nurse with specific training in the registry and its definitions\(^12\). Participation in this program has been shown to improve surgical outcomes and decrease post-operative morbidity\(^12\)\(^,\)\(^13\). Despite its usefulness, data collection remains labor intensive and participation is expensive\(^14\). With respect to patient functional health status, the NSQIP trained reviewers most often manually review charts to abstract a determination for the level of the patient’s functional health status placing patients into one of three categories (Table 1): independent, partially dependent, and totally dependent. An independent patient is defined as one who does not require assistance from another person for any activities of daily living, including one who functions independently with the use of prosthetics, equipment, and/or devices. A partially dependent patient requires some assistance from another person for activities of daily living regardless of use of prosthetics, equipment, and/or devices. Finally, a totally dependent patient requires total assistance for all activities of daily living. Functional health status in the
NSQIP database is determined within 30 days prior to the operation and is highly correlated with post-operative outcomes\textsuperscript{15}.

By reviewing and categorizing terms and phrases in clinical notes associated with functional health status, it may be possible to improve automated abstraction efforts. Functional health status is often poorly defined, or difficult to define with individual patients between different providers and scoring systems are not uniformly used. As a first step toward understanding the value of the clinical notes for functional health status, we sought to develop a library of terms associated with functional status. More broadly, our hypothesis is that much of the data relating to functional health status would be found within free-text documentation in clinical notes.

**Methods**

Institutional Review Board approval was obtained for this study. The study was completed at the University of Minnesota Medical Center, an integrated health system in partnership with Fairview Health Services based in Minneapolis, serving the upper Midwest. We performed our initial search using the medical center’s Clinical Data Repository (CDR). Data in from the CDR is compiled from the electronic health records of more than two million patients from eight hospitals and forty clinics\textsuperscript{16}. The CDR was queried for all patients included in the NSQIP database over three years (2013-2015). Of these patients, a stratified random convenience sample of 75 patients (twenty-five patients in each of the three functional health status categories as defined by NSQIP: “independent”, “partially dependent”, and “totally dependent”) were selected.

Physician reviewers (SS and EA) were blinded to the NSQIP functional health status determination for each patient. Reviewers examined all clinical notes and forms for the 30 days prior to the operative procedure for which functional health status was originally measured and determined. All phrases associated with functional health status were recorded. Details associated with the phrase, such as clinical note section, type of note, author credentials, and author specialty were also recorded.

After completing the chart review and recording all functional health status data, our reviewers assigned a NSQIP functional health status category and a Karnofsky Performance Score (Table 1) to each of the patient charts based only on functional health status terms recorded during the chart review. These scores were compared with the gold standard scores, which were the determinations previously made in the NSQIP registry by the SCR. To assess the inter-rater reliability of our functional health status determination of associated phrases, a subset of 8 overlapping patients (10.7\%) was performed by both reviewers with percentage agreement and Kappa calculated. Statistical analyses were performed using R software (Vienna, Austria, 2017).

**Table 1:** Karnofsky Performance Scale\textsuperscript{6,7} and NSQIP functional status scale

<table>
<thead>
<tr>
<th>Performance Score</th>
<th>Functional Status Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>No complaints, no evidence of disease</td>
</tr>
<tr>
<td>90</td>
<td>Able to complete major activities; minor signs and symptoms of disease</td>
</tr>
<tr>
<td>80</td>
<td>Normal activity with effort; some signs and symptoms of disease</td>
</tr>
<tr>
<td>70</td>
<td>Care of self; unable to carry on normal activities or do active work</td>
</tr>
<tr>
<td>60</td>
<td>Requires occasional assistance; able to care for most of personal needs</td>
</tr>
<tr>
<td>50</td>
<td>Requires considerable assistance and frequent medical care</td>
</tr>
<tr>
<td>40</td>
<td>Disabled; requires special care and assistance</td>
</tr>
<tr>
<td>30</td>
<td>Severely disabled; hospital admission is indicated; death not imminent</td>
</tr>
<tr>
<td>20</td>
<td>Very sick; hospital admission necessary and active treatment necessary</td>
</tr>
<tr>
<td>10</td>
<td>Moribund; fatal processes progressing</td>
</tr>
<tr>
<td>0</td>
<td>Death</td>
</tr>
</tbody>
</table>

**NSQIP functional status scale**

- **Independent**: Does not require assistance from another person for any activities of daily living, including one who functions independently with the use of prosthetics, equipment, and/or devices.
- **Partially Dependent**: Requires some assistance from another person for activities of daily living regardless of use of prosthetics, equipment, and/or devices.
- **Totally Dependent**: Requires total assistance for all activities of daily living.
Results
A total of 75 patient charts from 2013-2015 were reviewed. In these charts, a total of 1,353 clinical notes were reviewed. Within these clinical notes, there were 1,328 phrases identified which were associated with the determination of a patient’s functional health status. There was a good variety of surgical specialties represented by the operations for which functional status was assessed in the NSQIP registry. Given the interest in functional health status for this study, specialties that operate on problems associated with low functional health status are more highly represented (i.e., neurosurgery, plastic surgery, colorectal surgery, urology, orthopedic surgery).

Patient demographic and surgical information is summarized in Table 2.

Table 2: Patient Demographics and Surgical Specialty

<table>
<thead>
<tr>
<th></th>
<th>Median Age (Range)</th>
<th>Male Patients n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Patients</td>
<td>51.5 (21-91)</td>
<td>39 (52%)</td>
</tr>
<tr>
<td>NSQIP Functional Status</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>48 (21-77)</td>
<td>11 (44%)</td>
</tr>
<tr>
<td>Partially Dependent</td>
<td>59.8 (28-91)</td>
<td>13 (52%)</td>
</tr>
<tr>
<td>Totally Dependent</td>
<td>47.6 (22-75)</td>
<td>15 (60%)</td>
</tr>
<tr>
<td>Surgical Specialty</td>
<td>Charts n (%)</td>
<td></td>
</tr>
<tr>
<td>Urology</td>
<td>13 (17.3%)</td>
<td></td>
</tr>
<tr>
<td>Orthopedic Surgery</td>
<td>11 (14.7%)</td>
<td></td>
</tr>
<tr>
<td>Gynecologic Surgery</td>
<td>10 (13.3%)</td>
<td></td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>10 (13.3%)</td>
<td></td>
</tr>
<tr>
<td>General Surgery</td>
<td>8 (10.7%)</td>
<td></td>
</tr>
<tr>
<td>Colorectal Surgery</td>
<td>7 (9.3%)</td>
<td></td>
</tr>
<tr>
<td>Plastic Surgery</td>
<td>5 (6.7%)</td>
<td></td>
</tr>
<tr>
<td>Vascular Surgery</td>
<td>5 (6.7%)</td>
<td></td>
</tr>
<tr>
<td>Bariatric Surgery</td>
<td>3 (4%)</td>
<td></td>
</tr>
<tr>
<td>Otolaryngology</td>
<td>2 (2.7%)</td>
<td></td>
</tr>
<tr>
<td>Thoracic Surgery</td>
<td>2 (2.7%)</td>
<td></td>
</tr>
</tbody>
</table>

All of the 1,328 phrases were also annotated according to the type of clinical note in which they appeared as well as the clinical note section in which they appeared. These distinctions were evaluated separately. Breakdown of clinical note type and clinical note section can be found in Table 3. The “progress note” category within the clinical note type includes both daily progress notes written by physicians as well as progress and miscellaneous notes written by nursing and ancillary staff. In some cases, clinical note sections were not clearly delineated or present; these notes are included in the “not applicable” category. This study found that the majority of functional health status information was found in history & physical notes, anesthesia assessments, and office visits, which would be likely to have some component of assessment of patient functional status prior to an operation. The clinical note sections that featured the most functional health status data were history of present illness, assessment/plan, review of systems, and physical exam. All of these sections are likely to describe symptoms, major medical problems, and impairments that affect the patient.

Functional health status-related phrases were recorded in the electronic medical record most frequently by physicians. An analysis of functional health status terms ordered by author role is shown in Table 3. The vast majority of functional health status phrases were recorded by providers (physicians, trainees, midlevel provider).
Table 3: Functional Health Status Phrases by Clinical Note Type, Note Section, and Author

<table>
<thead>
<tr>
<th>Clinical Note Type</th>
<th>Phrases n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>History &amp; Physical</td>
<td>440 (33.1%)</td>
</tr>
<tr>
<td>Anesthesia Pre-Operative Assessment</td>
<td>338 (25.5%)</td>
</tr>
<tr>
<td>Office Visit</td>
<td>237 (17.8%)</td>
</tr>
<tr>
<td>Progress Note</td>
<td>160 (12.0%)</td>
</tr>
<tr>
<td>Consultation Note</td>
<td>69 (5.2%)</td>
</tr>
<tr>
<td>Emergency Department Visit</td>
<td>51 (3.8%)</td>
</tr>
<tr>
<td>Telephone Note</td>
<td>23 (1.7%)</td>
</tr>
<tr>
<td>Operative Note</td>
<td>8 (0.6%)</td>
</tr>
<tr>
<td>Discharge Summary</td>
<td>2 (0.2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical Note Section</th>
<th>Phrases n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of Present Illness</td>
<td>327 (24.6%)</td>
</tr>
<tr>
<td>Not Applicable</td>
<td>215 (16.2%)</td>
</tr>
<tr>
<td>Assessment/Plan</td>
<td>199 (15.0%)</td>
</tr>
<tr>
<td>Review of Systems</td>
<td>185 (13.9%)</td>
</tr>
<tr>
<td>Physical Exam</td>
<td>156 (11.7%)</td>
</tr>
<tr>
<td>Past Medical History</td>
<td>141 (10.6%)</td>
</tr>
<tr>
<td>Social History</td>
<td>38 (2.9%)</td>
</tr>
<tr>
<td>Past Surgical History</td>
<td>26 (2.0%)</td>
</tr>
<tr>
<td>Chief Complaint</td>
<td>18 (1.4%)</td>
</tr>
<tr>
<td>Form Elements</td>
<td>14 (1.1%)</td>
</tr>
<tr>
<td>Operative Indications</td>
<td>9 (0.7%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Author</th>
<th>Phrases n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff Physician</td>
<td>795 (59.9%)</td>
</tr>
<tr>
<td>Midlevel Provider</td>
<td>182 (13.7%)</td>
</tr>
<tr>
<td>Resident or Fellow</td>
<td>144 (10.8%)</td>
</tr>
<tr>
<td>Registered Nurse</td>
<td>64 (4.8%)</td>
</tr>
<tr>
<td>Other</td>
<td>30 (2.3%)</td>
</tr>
<tr>
<td>Physical Therapist</td>
<td>28 (2.1%)</td>
</tr>
<tr>
<td>Wound Ostomy Continence Nurse</td>
<td>26 (2.0%)</td>
</tr>
<tr>
<td>Occupational Therapist</td>
<td>25 (1.9%)</td>
</tr>
<tr>
<td>Medical Assistant</td>
<td>21 (1.6%)</td>
</tr>
<tr>
<td>Social Worker</td>
<td>18 (1.4%)</td>
</tr>
<tr>
<td>Medical Student</td>
<td>5 (0.4%)</td>
</tr>
</tbody>
</table>

Author specialty was also recorded. Functional health status data was recorded most frequently by anesthesiologists (23.6% of phrases) and internists (19.7% of phrases). When combined, surgical specialties amounted to 21.7% of the phrases and medical specialties recorded 34.3% of the phrases. Nursing and ancillary staff accounted for 19.7% of functional health status phrases.

Phrases related to functional health status were categorized into seven major categories including: diagnosis, activity/care needs, physical exam elements, functional scores, assistive equipment, symptoms, and surgical history. The phrases are divided according to category and NSQIP functional status determination in Table 4. The amount and proportion of functional health status-related diagnoses increased with increasing level of dependence.
Table 4: Categorized Phrases According to NSQIP Functional Category

<table>
<thead>
<tr>
<th>Phrase Category</th>
<th>Independent n (%)</th>
<th>Partial Dependence n (%)</th>
<th>Total Dependence n (%)</th>
<th>All Patients n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Phrases</td>
<td>209</td>
<td>607</td>
<td>512</td>
<td>1328</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>23 (11.0%)</td>
<td>194 (32.0%)</td>
<td>255 (49.6%)</td>
<td>472 (35.5%)</td>
</tr>
<tr>
<td>Activity/Care Needs</td>
<td>61 (29.2%)</td>
<td>161 (26.5%)</td>
<td>77 (15.0%)</td>
<td>297 (22.3%)</td>
</tr>
<tr>
<td>Physical Exam Elements</td>
<td>35 (16.7%)</td>
<td>71 (11.7%)</td>
<td>48 (9.3%)</td>
<td>154 (11.6%)</td>
</tr>
<tr>
<td>Functional Scores</td>
<td>47 (22.5%)</td>
<td>41 (6.8%)</td>
<td>32 (6.2%)</td>
<td>120 (9.0%)</td>
</tr>
<tr>
<td>Assistive Equipment</td>
<td>7 (3.3%)</td>
<td>65 (10.7%)</td>
<td>38 (7.4%)</td>
<td>110 (8.3%)</td>
</tr>
<tr>
<td>Symptoms</td>
<td>23 (11.0%)</td>
<td>45 (7.4%)</td>
<td>27 (5.3%)</td>
<td>95 (7.2%)</td>
</tr>
<tr>
<td>Surgical History</td>
<td>13 (6.2%)</td>
<td>30 (4.9%)</td>
<td>37 (7.2%)</td>
<td>80 (6.0%)</td>
</tr>
</tbody>
</table>

Unique phrases and terms were isolated from all phrases that were recorded during the chart review process. There was a total of 47 unique diagnoses. Unique diagnoses are listed in Table 5. Some of the diagnoses had associated modifiers, which were usually markers of severity or location. For example, the diagnosis “multiple sclerosis” had modifiers “with worsening plaques”, “progressive”, and “unclear”. “Pressure ulcer” had modifiers: “sacral”, “gluteal”, “coccygeal”, and “Stage IV” found in the clinical notes.

Table 5: Unique Diagnoses

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>scoliosis</td>
<td>kyphosis</td>
<td>lumbar stenosis</td>
<td>subdural hemorrhage</td>
<td>chronic pain</td>
</tr>
<tr>
<td>meningioma</td>
<td>CNS lymphoma</td>
<td>hemiplegia</td>
<td>monoplegia</td>
<td>autism</td>
</tr>
<tr>
<td>ulcer</td>
<td>neurogenic bowel</td>
<td>hip fracture</td>
<td>dementia</td>
<td>malnutrition</td>
</tr>
<tr>
<td>paralysis</td>
<td>Polio</td>
<td>post-Polio syndromes</td>
<td>hyper-reflexia</td>
<td>spasticity</td>
</tr>
<tr>
<td>spasmodic dysphonia</td>
<td>limb hypogenesis</td>
<td>congenital deformity</td>
<td>radiculitis</td>
<td>Cauda Equina</td>
</tr>
<tr>
<td>Alzheimer Disease</td>
<td>Parkinson Disease</td>
<td>mental retardation</td>
<td>Cerebral Palsy</td>
<td>cognitive defects</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>dysreflexia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>developmental delay</td>
</tr>
</tbody>
</table>

A large portion of patient activity level and care needs were unique; however, the themes of these phrases were similar. Usually, activities measured were similar but there were varying levels of dependence and assistance required between patients. Care needs varied slightly by level of care and nomenclature for facility type. Activity level was often judged based on several activities of daily living: ambulation/walking, eating/cooking, climbing stairs, transferring, toileting, bathing/hygiene, dressing, and exercise. General statements were sometimes made to summarize the patient’s level of activity, such as “not frail”, “good mobility”, or “low exercise capacity”. Care needs were relayed through assistive facility (nursing home, long term care facility, adult foster care, assisted living) or by the individuals helping with daily cares (husband caregiver, personal care assistant, home nurse, daily skilled nursing care).

Physical exam elements were classified into unique terms and phrases. There were 87 unique physical exam elements that could be further divided into three physical exam categories. There were 37 unique “motor/strength/sensation” items, 37 unique “general exam/appearance” items, and 13 unique “cognitive” items.

Functional status scoring systems were particularly useful for making a functional status determination, but were not present for all patients. These scoring systems included American Society of Anesthesiologists (ASA) Class, Karnofsky Functional Scale, Berg Balance Scale, METs estimation, the EQ-5D quality of life questionnaire, and the PROMIS questionnaire. METS and ASA were most prevalent in this chart review and frequently recorded by anesthesiologists, likely reflecting the peri-operative status of these patients.
Unique assistive equipment is recorded in Table 6. There were 31 unique equipment items/devices that were relevant to functional health status.

**Table 6: Unique Assistive Equipment**

<table>
<thead>
<tr>
<th>Spinal cord stimulator</th>
<th>Indwelling Foley</th>
<th>Wheelchair</th>
<th>Motorized wheelchair</th>
<th>Baclofen pump</th>
<th>Urinary catheters</th>
<th>Home oxygen</th>
<th>Lift chair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walker</td>
<td>Bath bench</td>
<td>Leg brace</td>
<td>Foot brace</td>
<td>Knee brace</td>
<td>Torso brace</td>
<td>Neck oxygen</td>
<td>Crutches</td>
</tr>
<tr>
<td>Scooter</td>
<td>Intrathecal pump</td>
<td>Home ramp</td>
<td>Prosthesis</td>
<td>Ostomy pouch</td>
<td>Jay cushion</td>
<td>Roho cushion</td>
<td>Hoyer lift</td>
</tr>
<tr>
<td>Shower chair</td>
<td>Cane</td>
<td>Bipap</td>
<td>Hospital bed</td>
<td>Stretcher</td>
<td>Grab bars</td>
<td>Ventilator</td>
<td></td>
</tr>
</tbody>
</table>

Twenty-eight unique symptoms were found in this study. Like diagnoses, modifiers of the symptoms typically highlighted symptom severity, frequency, and location. The unique symptoms related to functional health status determination are found in Table 7.

**Table 7: Unique Symptoms**

<table>
<thead>
<tr>
<th>Pain</th>
<th>Weakness</th>
<th>Shortness of breath</th>
<th>Tingling</th>
<th>Numbness</th>
<th>Spasticity</th>
<th>Fatigue</th>
<th>Secretion problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple falls</td>
<td>Hematuria with cath</td>
<td>Urinary incontinence</td>
<td>Fecal incontinence</td>
<td>Altered sensation</td>
<td>Paresis</td>
<td>Paresthesias</td>
<td>Radiculopathy</td>
</tr>
<tr>
<td>Worsening gait</td>
<td>Swelling</td>
<td>Urinary retention</td>
<td>Neuropathy</td>
<td>Worsening motor function</td>
<td>Worsening neurologic status</td>
<td>Seizures</td>
<td>Unresponsive</td>
</tr>
<tr>
<td>Combative behavior</td>
<td>Constant movement</td>
<td>Memory deficit</td>
<td>Slurred speech</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Finally, there were terms related to surgical procedures that were helpful for functional status determinations. There were 22 unique phrases in the surgical history that helped with determination listed in Table 8.

**Table 8: Unique Surgical Terms**

<table>
<thead>
<tr>
<th>Epidural injection</th>
<th>Below knee amputation</th>
<th>Thoracic spine surgery</th>
<th>Artificial urinary sphincter</th>
<th>Suprapubic catheter placement</th>
<th>Mitrofanoff</th>
<th>Ventriculoperitoneal shunt</th>
<th>Ventriculopleural shunt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neck fusion</td>
<td>Above knee amputation</td>
<td>Tracheostomy</td>
<td>Colostomy</td>
<td>Craniotomy</td>
<td>Urinary diversion</td>
<td>Nephrostomy tubes</td>
<td>Percutaneous gastrostomy</td>
</tr>
<tr>
<td>Ileal conduit</td>
<td>Monti</td>
<td>Bladder augmentation</td>
<td>Urostomy</td>
<td>Disarticulation</td>
<td>Gastro-jejunostomy</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sensitivity was measured for comparing functional status designation using only functional status level phrases identified in the review and NSQIP surgical clinical reviewer determination. Sensitivity decreased as functional status complexity increased. A sensitivity of 96% was obtained for independent patient identification, 60% for identification of partially dependent patients, and 44% for totally dependent patients. Table 9 demonstrates the distribution of the designation of functional status based on the NSQIP SCR designation as the gold standard and functional status level phrases in separate human reviewer designation.
Table 9: NSQIP Functional Status (Reviewer Designation versus Gold Standard)

<table>
<thead>
<tr>
<th>SCR Designation (gold standard)</th>
<th>Independent</th>
<th>Partially Dependent</th>
<th>Totally Dependent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Reviewer Designation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Independent</td>
<td>24</td>
<td>6</td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td>Partially Dependent</td>
<td>1</td>
<td>15</td>
<td>13</td>
<td>29</td>
</tr>
<tr>
<td>Totally Dependent</td>
<td>0</td>
<td>4</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>75</td>
</tr>
</tbody>
</table>

\[ p = 3.2201860600803914 \times 10^{-11} \]

The gold standard functional health status designation was also compared against the Karnofsky Performance Scale as determined by independent review (Figure 1). Independent status had a maximum score of 100, median of 90, and minimum of 70 (interquartile range [IQR] 80-100). Partially dependent status had a maximum score of 100, median of 60, and a minimum of 40 (IQR 60-70). Totally dependent patients had a maximum score of 70, median of 50 and minimum of 30 (IQR 40-50).

Figure 1: Karnofsky Score versus NSQIP Functional Status

Box Plot of Karnofsky Score Compared to NSQIP Functional Status Categories

Lastly, phrases were assessed for their relevance to functional health status by a second reviewer (EA). Ten percent of the phrases were used for assessment. Inter-rater reliability for determination if a phrase was important for functional health status was scored with an agreement of 90.7% and a Cohen’s kappa of 0.737.

Discussion

At present, functional health status data has not been well integrated into electronic medical records and most clinical workflows in practice. The combination of use of a multitude of functional health status assessments along with using clinical judgement and incomplete/inconsistent documentation make this integration challenging. The purpose of this study was to characterize signals for functional health status in clinical notes to attempt to organize and classify functional health status-related data. This study demonstrated that a variety of phrase categories can be helpful for determining functional health status including diagnoses related to functional health status, activity descriptions, home care needs, physical exam, functional scores, assistive equipment and medical devices, symptoms, and surgical procedures. Anesthesiologists and internal medicine physicians were the most frequent recorders of functional health status data in this study, likely because of the perioperative status of the patient.
There were a large number of unique terms for patient activity level and home care needs. The remaining categories had a relatively small number of unique phrases, but many modifiers related to severity, frequency, and location (particularly for diagnoses and symptoms).

We found that correctly identifying NSQIP functional health status category (gold-standard) with functional health status-related terms was more challenging in our chart review with patients of increasing functional health status complexity. As seen in Table 9, there was still correlation between outcomes, however, there was particularly increased variability in the designation for the “totally dependent” category. This could be because functional health status is inherently difficult to classify. Alternatively, it is often difficult to make a determination based on descriptions of a complex topic such as functional health status which can span many different observation categories, particularly when different providers’ descriptions may not align. There were two cases that were classified as “totally dependent” by the SCR, yet “independent” by human reviewer. Because of the wide discrepancy, these cases were re-reviewed and found to have conflicting data in clinical notes. In these patients, there were signals of total/partial dependency for functional health status in some notes, while other providers’ notes clearly stated that the patients were completely independent. This highlights the complexity of functional health status and limitations associated with designating functional health status retrospectively.

This study is limited in its retrospective nature. Also, there are likely additional factors that were involved in the determination of functional health status level by the physicians performing the assessment that were not entered in the electronic medical record. As such, determining functional status in a prospective manner in a standard and rigorous fashion is most ideal. Additionally, this is a single institution study with a study of relatively small sample size, and the description of functional health status may differ between institutions or even between surgical services. In our cohort, several author types (physical therapist, occupational therapists, and social workers) may have been particularly underrepresented. These author types wrote 1.2% of the total clinical notes reviewed, yet contributed to 5.3% of the functional health status-related phrases. It is likely that these authors have a larger impact on the functional health status descriptors than represented in this study. This study reviewed surgical patients, which may have contributed to the lack of documentation from these authors. Perhaps the reason for low percentages of these notes overall is that for cases where patients were totally independent or totally dependent (n=50), physical therapy, occupational therapy, and social work may have been minimally involved, since the supportive needs were either very low, or conversely, maximal and established. Finally, certain surgical services more involved in the care of lower functional status patients were likely over-represented.

Other methods have been used to attempt to characterize functional health status. Another retrospective review was performed which collected functional health status terms within the Veteran Affairs electronic medical record and patient-reported functional health status data on social media, concluding that standard terminologies such as Unified Medical Language System do not sufficiently cover functional health status information. In a separate project, the same group used topic modeling as a method to extract relevant frailty information in clinical text. A study by Ruggieri et al. found that functional health status terminologies centered around verb phrases, particularly descriptions of motion. A “frame-semantic” method was used for functional health status representation with a final goal of improving information abstraction and natural language applications of functional health status.

Similar work characterizing clinical note data has been performed at our institution to develop an automated method of data abstraction for determination of surgical site infections and other complications. Using keywords related to these complications improved the accuracy of the detection of these complications. We identified and categorized terms and phrases used in surgical site infection descriptions in a previous study. Translating this method to this current project, the new terms identified in this study could be valuable for the automated detection of functional health status level. Additionally, these terms could potentially be incorporated into an automated approach to examine notes as they are being written to determine if they meet minimum documentation requirements for functional health status. We believe that expansion of our previous work in an analogous fashion can improve automated detection techniques for functional health status determinations.

**Conclusion**

Functional health status is a difficult clinical entity to quantify. Determination of level of functional status likely differs between providers, and while functional status scores are often helpful in this determination, they are not always documented. Factors in functional health status determination can be found in clinical notes through diagnoses, activity and home care descriptions, physical exam elements, functional scores, assistive equipment, symptoms, and surgical procedures. The phrases identified in this study could potentially be used to assist in automation of detection of functional health status level.
Acknowledgements

This research was supported by the University of Minnesota Academic Health Center Faculty Development Award, Agency for Healthcare Research and Quality (R01HS24532), National Institutes of Health (NIH) Clinical and Translational Science Award (CTSA) program (UL1TR000114), NIH/National Institute of General Medical Sciences (NIGMS) (R01GM120079), Fairview Health Services, and University of Minnesota Physicians.

References


A Novel Representation of Vaccine Efficacy Trial Datasets for Use in Computer Simulation of Vaccination Policy

Mohammadamin Tajgardoon\textsuperscript{1}, Michael M. Wagner\textsuperscript{1,2}, Shyam Visweswaran\textsuperscript{1,2}, Richard K. Zimmerman\textsuperscript{3}

\textsuperscript{1}Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA; \textsuperscript{2}Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA; \textsuperscript{3}Department of Family Medicine, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Computer simulation is the only method available for evaluating vaccination policy for rare diseases or emergency use of new vaccines. The most realistic simulation of vaccination policy is agent-based simulation (ABS) in which agents have similar socio-demographic characteristics to a population of interest. Currently, analysts use published information about vaccine efficacy (VE) as the probability that a vaccinated agent develops immunity; however, VE trials typically report only a single overall VE, or VE conditioned on one covariate (e.g., age). Thus, ABS’s potential to realistically simulate the effects of co-existing diseases, gender, and other characteristics of a population is under-used. We developed a Bayesian network (BN) model as a compact representation of a VE trial dataset for use in ABS of vaccination policy. We compared BN-based VEs to the VEs estimated directly from the dataset. Our evaluation results suggest that VE trials should release statistical models of their datasets for use in ABS of vaccination policy.

1 Introduction

Computer simulation is an important method that allows policy makers to examine the effectiveness of vaccination strategies in a virtual environment before applying them to a real population\textsuperscript{1}. It is really the only method available for rare diseases (e.g., study of smallpox vaccination\textsuperscript{2}), or emergency use of new vaccines (e.g., study of vaccination policy for Ebola and Zika epidemics\textsuperscript{3–5}).

Agent-based simulation (ABS) is the most realistic method for computer simulation of vaccination policy. Briefly, when using an ABS to analyze a vaccination policy $P$, analysts first create a population of agents whose socio-demographic and behavioral characteristics are similar to a population of interest. They then program the simulator to emulate policy $P$ and emergence of an epidemic. In particular, analysts use published data about vaccine efficacy (VE) to determine the probability of disease transmission to a vaccinated agent.

There is usually limited information about VE available to ABSs of vaccination policy since VE trials typically report only a single overall VE, or VE conditioned on only one covariate (e.g., age)\textsuperscript{6–8}. Additionally, while it is possible to recover other conditional VEs from VE trial datasets, these datasets are not released to the public domain and they are difficult to obtain even under restrictions. Thus, ABSs of vaccination policy are not able to simulate the effects of co-existing diseases, medications, gender, and other socio-demographic characteristics of a population.

Therefore, our objective is to improve the information available about VE. In particular, we developed and evaluated a Bayesian network (BN) model as a compact representation of a VE trial dataset for use in an ABS of vaccination policy. A BN is a probabilistic model that encodes a joint probability distribution over all dataset variables. VE conditioned on any set of covariates can be estimated from a BN model of a VE trial dataset. We discuss the details of estimating VE from a BN model in Section 3.

The VE trial dataset that we used was collected in a study of a trivalent inactivated influenza vaccine (TIV), manufactured by GlaxoSmithKline (GSK) Biologicals, during the 2006–2007 influenza season in the Czech Republic and Finland\textsuperscript{8}. We used two score-based BN learning algorithms in combination with multiple score functions to develop several BN models of the VE trial dataset, in addition to a naïve Bayes model as the baseline model. We evaluated the models in 10-fold cross validations to find the model that fits the dataset the best. We compared the VEs estimated from the best BN model to the VEs that we directly estimated from the dataset. Our results suggest that VE trials consider releasing a BN model of their datasets which contains significantly more information about VE compared to the tables they report in their publications.
2 Related Work

Previous studies have addressed the problem that a single VE may not apply to all individuals\textsuperscript{9–15}. The issue is usually referred to as heterogeneity in VE and has been investigated since 1915 when Greenwood and Yule\textsuperscript{9} noticed the problem. Longini et al.\textsuperscript{13} categorized the approaches to modeling the heterogeneity in VE into two groups: (1) by using stratification\textsuperscript{10,12}, and (2) by allowing the susceptibility in vaccinated persons to follow a probability distribution\textsuperscript{11,13,16}. In the stratification approach, vaccinated individuals are stratified into a number of mutually exclusive groups and VE is estimated for each group. The number of strata in this approach is usually determined manually and depends on whether the heterogeneity in VE originates from the host characteristics or is vaccine-related\textsuperscript{12}. In the second approach, the vaccine effect on individuals follows a probability distribution. For example, Longini et al.\textsuperscript{13} defined a parameter, $\alpha$, as a fraction of the vaccinated persons who may develop complete protection. The remaining fraction, $1 - \alpha$, may receive only partial protection. Their objective was to estimate $\alpha$ and the parameters for the distribution of individuals with partial protection from trial or observational data. To model VE, they used a frailty mixture model which is a class of survival models that allow some vaccinated individuals to be more susceptible than other vaccinees.

Moreover, in a recent study, Mehtäliä et al.\textsuperscript{15} investigated how to estimate the efficacy of a vaccine with heterogeneous protection in a SIS (Susceptible-Infected-Susceptible) model. They modeled VE using three parameters: the proportion of vaccinated completely protected from infection, the relative attack rates in acquisition phase, and the relative attack rates in clearance phase, for a vaccinated person not completely protected as compared to an unvaccinated individual.

In this paper we propose a different approach for modeling the vaccine effect in the presence of heterogeneity, which to our knowledge has not been tried previously. We use machine-learning algorithms to fit a joint probability distribution over all variables in a vaccination trial dataset including socio-demographic and disease characteristics, and vaccination-related covariates. We expect that the machine-learning algorithms will automatically capture the host or vaccine-related heterogeneity in the vaccine effect, as long as the relevant covariates are measured in the vaccination trial.

3 Method

We propose a machine-learning based method to fit a joint probability distribution over variables in a VE trial dataset and then use the distribution to estimate VE conditioned on any set of covariates. In the following sections, we first describe BNs and the learning algorithms we use. Then, we explain how to estimate VE from a BN model of a VE trial dataset.

3.1 Bayesian Networks (BN)

A BN over a set of random variables $V$ is a pair $BN = (G, P)$, where $G$ is a directed acyclic graph (DAG) and $P$ is a joint probability distribution of $V$. Each node in the DAG is assigned a conditional probability distribution as a function of the node’s parents. Nodes with no parents have a marginal probability distribution. Arcs represent dependency relationships between variable pairs\textsuperscript{17}. Learning a BN model of a dataset includes two steps: (1) DAG construction, in which a BN structure-learning algorithm is used to find a DAG from data, and (2) parameter estimation, which is estimating the probability distribution of nodes in the DAG.

We used two score-based structure-learning methods: hill-climbing (HC) with random restart and tabu search. These algorithms start from an initial network structure (e.g., an empty graph) and add, delete, and reverse arcs until a score function is no longer improved. Tabu search algorithm maintains a list of recent operations (i.e., tabu list) to avoid repeating the most recent actions and becoming stuck in a suboptimal solution\textsuperscript{18}. We tested several score functions that are common in the literature: Bayesian information criterion (BIC)\textsuperscript{19}, Bayesian Dirichlet equivalence uniform (BDeu)\textsuperscript{17}, and K2\textsuperscript{20}.

Once a DAG is learned by a BN structure-learning algorithm, a parameter estimation method is used to learn the conditional probability distributions of the nodes given the DAG and data. We use the maximum a posteriori (MAP) parameter estimation method in our analysis.
After learning a BN model from data, i.e., learning the structure and estimating the parameters, we use inference algorithms to obtain the probability of an event of interest (e.g., disease contraction) conditioned on a set of evidence (e.g., vaccinated and female). This process is called inference or probabilistic reasoning. BN inference algorithms are categorized into exact and approximate inference algorithms. Exact inference algorithms use Bayes theorem and Markov condition to derive exact values of conditional probabilities and are limited to small BNs. Approximate inference algorithms use Monte Carlo sampling to estimate conditional probabilities from a large set of dataset instances generated from the BN. These algorithms perform better on large BNs. In our experiments we use logic sampling which is an approximate inference algorithm.

3.2 Estimate VE from a BN Model

In this section, we explain how to estimate VE from a joint probability distribution that is obtained by a BN learning algorithm. In particular, we formulate VE as a function of two conditional probabilities that are inferred from a joint probability distribution.

VE is measured as the proportional reduction in disease incidence in vaccinated group compared to unvaccinated group. When we do not condition on any other covariates, e.g., age, gender, etc., VE is referred to as Overall VE and is calculated using the following equation:

\[ VE = \frac{ARU - ARV}{ARU} = 1 - \frac{ARV}{ARU} \]  

where ARV and ARU are attack rates in the vaccinated and unvaccinated groups, respectively. According to the relative frequency definition of probability, ARV can be defined as the probability of disease incidence in vaccinees, i.e., Pr(disease|vaccinated), and similarly, we can formulate ARU as the probability of contacting the disease in non-vaccinees, i.e., Pr(disease|unvaccinated). Therefore, by replacing attack rates with probabilities in Equation 1, we can re-write VE as a function of two conditional probabilities:

\[ VE = 1 - \frac{Pr(disease|vaccinated)}{Pr(disease|unvaccinated)} \]  

We can extend Equation 2 to obtain VE conditioned on a set of covariates (e.g., age, gender). Let \( X = \{X_1, X_2, ..., X_n\} \) be a set of covariates in a VE trial dataset, which are initialized with values \( x = \{x_1, x_2, ..., x_n\} \). VE conditioned on \( X = x \) is formulated as follows:

\[ [VE|X = x] = 1 - \frac{Pr(disease|vaccinated, X = x)}{Pr(disease|unvaccinated, X = x)} \]  

4 Dataset Description

The dataset that we used to develop and evaluate BN models was acquired in a randomized, double-blind, placebo-controlled study of a trivalent inactivated split-virus influenza vaccine (TIV) manufactured by GlaxoSmithKline (GSK) Biologicals. The study was conducted during the 2006–2007 influenza season in the Czech Republic and Finland. The primary objective of the study was to assess the efficacy of the TIV in the prevention of culture-confirmed influenza due to strains antigenically matched to the vaccine.

The study subjects were 7652 healthy adults between 18 and 64 years old. 5103 of the subjects were vaccinated with 1 dose of TIV and the remaining 2549 received 1 dose of placebo in their first visit. Table 1 demonstrates the demographics of vaccine and placebo groups. Due to the randomization in group assignments, the distribution of demographic characteristics is similar in both groups.

The dataset contains 38 tables which fall into two categories of administrative (30 tables) and analysis (8 tables). Administrative tables are populated with the study meta-data, including the list of codes for activities and events, information on dataset fields, and administrative details (e.g., consent forms, eligibility and elimination codes, protocol
violations). The analysis tables contain the information related to vaccination, lab results, demographic characteristics, and medical conditions (Table 2).

We used only variables from six analysis tables. We excluded the two remaining analysis tables, influenza-like-illness (ILI) episodes and serology information, for two reasons: (1) ILI variables, e.g., symptoms and medications during ILI episodes, were post-incidence events that could not influence the VE, and (2) serology tests were performed for only a subset of participants in the study. The outcome variable that we used for model evaluations and VE estimations is influenza due to vaccine antigenically matched strains, which is the same outcome variable that the VE trial used for their primary analysis.

For our analysis we created a single table with 133 variables including demographics, medical conditions, concomitant vaccinations, non-ILI-episode medications, vaccine-related variables, and the outcome variable (influenza due to vaccine antigenically matched strains). All variables were either binary or categorical with a few categories except for Age which had integer values between 18 and 64. We discretized the Age variable to reduce its number of possible values. Specifically, at each iteration of a 10-fold cross-validation we applied a supervised discretization method to categorize the Age variable into a few bins.

Table 1: Demographic characteristics of the VE trial subjects. Second and third columns list the number and percentage of each demographic characteristic within vaccine and placebo groups, respectively. For Age, the average age and the range for each group is listed.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Vaccine (n=5103)</th>
<th>Placebo (n=2549)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean years (range)</td>
<td>39.94 (18-64)</td>
<td>39.74 (18-64)</td>
</tr>
<tr>
<td>Sex, no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>3069 (60.14)</td>
<td>1542 (60.49)</td>
</tr>
<tr>
<td>Race, no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>African Heritage or African American</td>
<td>1 (0.02)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Asian - East Asian</td>
<td>1 (0.02)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>White - Arabic or North African</td>
<td>3 (0.06)</td>
<td>1 (0.02)</td>
</tr>
<tr>
<td>White - Caucasian or European</td>
<td>5097 (99.88)</td>
<td>2547 (99.92)</td>
</tr>
<tr>
<td>Other</td>
<td>1 (0.02)</td>
<td>1 (0.04)</td>
</tr>
<tr>
<td>Ethnicity, no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>American Hispanic or Latino</td>
<td>5 (0.10)</td>
<td>1 (0.04)</td>
</tr>
<tr>
<td>Not American Hispanic or Latino</td>
<td>5098 (99.90)</td>
<td>2548 (99.96)</td>
</tr>
<tr>
<td>Country, no. (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech Republic</td>
<td>2664 (52.20)</td>
<td>1332 (52.26)</td>
</tr>
<tr>
<td>Finland</td>
<td>2439 (47.80)</td>
<td>1217 (47.74)</td>
</tr>
</tbody>
</table>

Table 2: List of 8 analysis tables and their variables. The VE trial dataset contained 38 tables including 30 administrative and 8 analysis tables. We used only the analysis tables to create a single table with 133 variables.

<table>
<thead>
<tr>
<th>Table name</th>
<th>Variable names (possible values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>age, gender, race, ethnicity, country, vaccination center</td>
</tr>
<tr>
<td>Medical condition</td>
<td>diagnosis, status (past, current, both)</td>
</tr>
<tr>
<td>Concomitant vaccination</td>
<td>vaccine name, code, vaccination date</td>
</tr>
<tr>
<td>Medication</td>
<td>name, code, start date, end date, daily dose</td>
</tr>
<tr>
<td>Vaccine</td>
<td>vaccine (Fluarix, placebo), vaccine vial-type (original, replacement), vaccination route, vaccine dose</td>
</tr>
<tr>
<td>Lab results</td>
<td>influenza type (A, B), virus strain (H1N1, H3N2, B), influenza due to vaccine antigenically matched strains (positive, negative), test type (HAI, QPCR)</td>
</tr>
<tr>
<td>Influenza-Like Illness (ILI) episodes</td>
<td>ILI number, start date, end date, symptoms, medications (for a subset of 504 subjects)</td>
</tr>
<tr>
<td>Serology results</td>
<td>seropositivity status</td>
</tr>
</tbody>
</table>
4.1 Bivariate Analysis of Covariates

We explored the statistical dependencies to the outcome variable using bivariate logistic regressions. In particular, for each variable \( X_i \) in the dataset, we fit a logistic regression model with the outcome as the dependent variable and \( X_i \) as the independent variable. Only a few variables demonstrated significant correlation to the outcome variable (i.e., p-value < 0.05) (Table 3). These variables are likely to have an arc to the outcome node in the BN structure of the dataset. Most of the correlated variables are vaccine-related (i.e., vaccine, vaccine lot, vaccine vial type, and ATP vaccination).

Table 3: List of variables with significant association with the outcome, influenza due to antigenically matched strains. For each variable in the dataset, we used a bivariate logistic regression with the outcome variable as the dependent variable. Here we list variables with coefficient p-value < 0.05 in their corresponding logistic regression model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>no. of influenza cases within each group (out of 124 total cases)</th>
<th>Estimated coefficient (CI. 95%)</th>
<th>Coefficient p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>age ( a )</td>
<td></td>
<td>-0.02 (-0.03 -0.01)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>country ( b )</td>
<td>81</td>
<td>0.55 (0.18 0.93)</td>
<td>0.003</td>
</tr>
<tr>
<td>vaccine ( c )</td>
<td>49</td>
<td>-1.14 (-1.5 -0.78)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>vaccine lot ( d )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lot1</td>
<td>25</td>
<td>-1.12 (-1.59 -0.67)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>lot2</td>
<td>24</td>
<td>-1.16 (-1.64 -0.71)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>vaccine vial type ( e )</td>
<td></td>
<td>-2.32 (-3.64 -0.46)</td>
<td>0.002</td>
</tr>
<tr>
<td>ATP vaccination ( f )</td>
<td></td>
<td>1.63 (-0.2 2.86)</td>
<td>0.027</td>
</tr>
</tbody>
</table>

\( a \) age was analyzed as a continuous variable.
\( b \) variable values are: 0=Finland (FI), 1=Czech Republic (CZ)
\( c \) variable values are: 0=placebo, 1=Fluarix vaccine
\( d \) variable values are: "placebo", "lot1", "lot2"
\( e \) variable values are: 0= replacement vaccine (due to unusability of study vaccine), 1=study vaccine
\( f \) variable values are: 0= vaccine not administered According To Protocol (ATP), i.e., vaccine received in the dominant arm or a replacement vial was administered, 1=vaccine administered ATP

5 Evaluation Method

5.1 Model Evaluation

To find the best BN model of the dataset, we performed model selection among several BN models. In particular, we evaluated the goodness of fit and the probability errors of the models in 10-fold cross validations. The following sections introduce the evaluation metrics.

5.1.1 Goodness of Fit Test

We evaluated the goodness of fit of BN models by two statistical tests which we call global test and local test\(^{24}\). The global test examines the null hypothesis that the observed data instances are occurring with the probabilities stated by the model. We constructed a test statistic using a logarithmic score as follows. Let \( Y \) be a binary outcome variable, \( y_i \) be the observed value of \( Y_i \) for instance \( i \), and \( p_i(Y_i = y_i) \) be the probability of \( Y_i = y_i \) according to the model. Then, a logarithmic score of instance \( i \) is:

\[
S_i = - \log p_i(Y_i = y_i).
\]

For \( N \) data instances the cumulative logarithmic score \( S \) would be the sum of the logarithmic scores over \( N \) instances, i.e.,

\[
S = \sum_{i=1}^{N} S_i = \log \prod_{i=1}^{N} p_i(Y_i = y_i). \quad (4)
\]
We can use $S$ to construct a standardized test statistic $Z$:

$$Z = \frac{S - \sum_{i=1}^{N} E_i}{\sqrt{\sum_{i=1}^{N} V_i}}$$  \hspace{1cm} (5)$$

where

$$E_i = - \sum_{k \in \{0, 1\}} p_i(Y_i = k) \log p_i(Y_i = k),$$  \hspace{1cm} (6)$$

$$V_i = \sum_{k \in \{0, 1\}} p_i(Y_i = k) \log^2 p_i(Y_i = k) - E_i^2.$$  \hspace{1cm} (7)$$

For large $N$, $Z$ is approximately distributed as a standard normal under the null hypothesis that the model fits to the data. Therefore, we can calculate the probability of seeing values as extreme or more extreme than $Z$ as a measure of the fitness of the model.

The local goodness of fit test evaluates the conditional probability distribution of the outcome node in a BN model. Let $Y$ be a binary outcome variable, $Y_i$ be the outcome variable for instance $i$ in the data, $y_i$ be the observed value of $Y_i$, and $X_{pa_i} = \rho$ be the instantiation of parents of node $Y$ in the data instance $i$. The local score of the outcome node in a BN model is computed via Equation 8,

$$LS_i = - \log p_i(Y_i = y_i | X_{pa_i} = \rho)$$  \hspace{1cm} (8)$$

where $p_i(Y_i = y_i | X_{pa_i} = \rho)$ is the probability of $Y_i = y_i$ given the configuration of parents of node $Y$ learned in the parameter learning step without doing inference. The standardized local score is obtained similar to the global test by use of Equations 4, 6, 7, and 5, respectively.

### 5.1.2 Outcome Probability Error

We compare the BN models in 10-fold cross validations with respect to their accuracy in estimating the probability of the outcome. We use mean squared error (MSE) and two probability calibration error metrics. MSE is the average of squared differences between the estimated probabilities and the corresponding value of the outcome variable (i.e., contracting influenza due to vaccine antigenically matched strains $\in \{0, 1\}$), and is calculated using the following formula:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (e_i - y_i)^2$$  \hspace{1cm} (9)$$

where $e_i$ is the estimated probability of outcome for data instance $i$, $y_i$ is the actual value of outcome for data instance $i$, and $N$ is the total data instances.

Probability calibration error is the discrepancy between probabilities inferred from a model and probabilities obtained from the dataset. To compute this error, we sort the estimated probabilities in an increasing order and divide them into $k$ number of bins. The model is perfectly calibrated if the mean of estimated probabilities are equal to the actual fraction of positive outcomes in each bin. The difference between the mean of probabilities and the fraction of positives in each bin is the base for probability calibration error metrics. We used maximum calibration error (MCE) and expected calibration error (ECE) metrics. MCE (Equation 11) measures the maximum calibration error over all bins, and ECE (Equation 10) calculates the average calibration error over all bins:

$$ECE = \sum_{i=1}^{k} P(i) \cdot |o_i - e_i|$$  \hspace{1cm} (10)$$

$$MCE = \max_{i=1}^{k} (|o_i - e_i|)$$  \hspace{1cm} (11)$$

where $k$ is the number of bins, $o_i$ is the fraction of observed positive outcomes in bin $i$, and $e_i$ is the mean of the estimated probabilities in bin $i$. 

5.2 Evaluating VEs estimated from a BN Model

We evaluate the VEs estimated from the BN model by comparing them to the same VEs acquired directly from the dataset. To estimate a VE from the BN model, we obtain two conditional probabilities and use Equation 3 to calculate the VE. For example, to estimate VE for females, two conditional probabilities are inferred from the BN model: the probability of disease contraction given vaccinated and being female, \( \Pr(\text{disease}|\text{vaccinated}, \text{female}) \), and the probability of disease contraction given unvaccinated and being female, \( \Pr(\text{disease}|\text{unvaccinated}, \text{female}) \). Each VE estimation is repeated 100 times. The average and standard deviation (SD) of 100 VEs is used as the estimate and the uncertainty interval, respectively. For estimating a VE directly from the dataset, we use Risk ratio test and report a 95% confidence interval. P-values are calculated using two-sided Fisher’s exact test.

6 Evaluation Results

6.1 Model Evaluation Result

We performed two goodness of fit tests followed by probability error measurements. All evaluations were implemented in a 10-fold cross validation. Table 4 shows the result of goodness of fit tests for different BN models and naïve Bayes. According to the global test there is significant evidence that the naïve Bayes model does not fit to the dataset. However, there is no evidence that BN models misfit the dataset. The local test demonstrates that there is no significant disagreement between the conditional probability distribution of the outcome node and the observed distribution in the dataset. We did not repeat the local test for naïve Bayes model as its overall fitness was already rejected by the global test.

Table 5 demonstrates the outcome probability error measures computed in a 10-fold cross validation. mean square error (MSE), maximum calibration error (MCE), and expected calibration error (ECE) was calculated based on the probability of the outcome in each test instance. All BN models have low calibration errors compared to the naïve Bayes model. MSE is similar for all models.

Table 4: Goodness of fit test. We performed two tests to evaluate the models’ goodness of fit in 10-fold cross validations. In the global test, we tested the null hypothesis that the model fits to the data. The local test evaluates the null hypothesis that the probability distribution learned for the outcome variable fits to the data. None of the BN models was rejected according to the p-values, except for the baseline model (naïve Bayes). We did not perform the local test for the naïve Bayes model.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Score function</th>
<th>Global score</th>
<th>P-value</th>
<th>Local score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabu</td>
<td>BIC</td>
<td>-0.31</td>
<td>0.38</td>
<td>-0.09</td>
<td>0.22</td>
</tr>
<tr>
<td>Tabu</td>
<td>K2</td>
<td>0.16</td>
<td>0.44</td>
<td>0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>Tabu</td>
<td>BDeu</td>
<td>0.35</td>
<td>0.36</td>
<td>0.12</td>
<td>0.2</td>
</tr>
<tr>
<td>HC</td>
<td>BIC</td>
<td>-0.31</td>
<td>0.38</td>
<td>-0.09</td>
<td>0.22</td>
</tr>
<tr>
<td>HC</td>
<td>K2</td>
<td>0.16</td>
<td>0.44</td>
<td>0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>HC</td>
<td>BDeu</td>
<td>0.35</td>
<td>0.36</td>
<td>0.12</td>
<td>0.2</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>-</td>
<td>13.1</td>
<td>&lt;.001</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5: Probability error measures via 10-fold cross validation. For each model, mean square error (MSE), maximum calibration error (MCE), and expected calibration error (ECE) computed for the outcome variable.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Score function</th>
<th>MSE</th>
<th>MCE</th>
<th>ECE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tabu</td>
<td>BIC</td>
<td>0.126</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>Tabu</td>
<td>K2</td>
<td>0.127</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td>Tabu</td>
<td>BDeu</td>
<td>0.127</td>
<td>0.010</td>
<td>0.005</td>
</tr>
<tr>
<td>HC</td>
<td>BIC</td>
<td>0.126</td>
<td>0.009</td>
<td>0.003</td>
</tr>
<tr>
<td>HC</td>
<td>K2</td>
<td>0.127</td>
<td>0.008</td>
<td>0.003</td>
</tr>
<tr>
<td>HC</td>
<td>BDeu</td>
<td>0.127</td>
<td>0.010</td>
<td>0.005</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>-</td>
<td>0.141</td>
<td>0.072</td>
<td>0.015</td>
</tr>
</tbody>
</table>
6.1.1 Final BN Model

All BN learning algorithms performed equally well (except for naïve Bayes) according to the model evaluations. We decided to continue the analysis with the tabu search algorithm and BDeu score function. We used a model averaging technique to build the final BN model\textsuperscript{27,28}. In particular, we trained 100 BN models over 100 bootstrap samples of the dataset. Then, we built a single BN model from the 100 BNs by selecting the arcs with proportional frequency of greater that or equal to 0.5. Since the resulted BN’s structure was too large to fit in here, Figure 1 illustrates a sub-graph of the structure which includes the outcome node (influenza due to antigenically matched strains) and its first and second level neighbors.

![Figure 1: A sub-graph of the average BN structure learned by the tabu search algorithm. 100 instances of BN models learned from 100 bootstrap samples of the data. An average BN model was built from arcs with proportional frequency $\geq 0.5$. The illustrated sub-graph includes the outcome variable (influenza due to antigenically matched strains) and its first and second level adjacent variables. Three variables (age, vaccine, vaccine vial type) form the Markov blanket of the outcome variable. The outcome variable is independent of other variables given its Markov blanket\textsuperscript{21}. * symbol in the figure indicates a diagnosis variable.]

6.2 Estimate VE from the BN Model

We estimated a number of VEs from both the final BN model and the dataset. Table 6 compares VEs estimated from the BN model to VEs directly estimated from the dataset. VEs obtained from the two approaches are the same or very close to each other. The uncertainty intervals for the BN-based estimates are also relatively similar to the confidence intervals of the direct estimates, i.e., when the confidence intervals become wider, the uncertainty intervals follow the same pattern (see the last row in Table 6).

7 Discussion and Conclusion

We developed and evaluated BN models as more complete representations of VE trial datasets compared to tables in the publications reported by VE trials. We performed model evaluations using the goodness of fit tests and probability error metrics to select the best BN model. We compared the VEs estimated from the BN model to the VEs estimated directly from the dataset. The results were promising as the BN-based estimates were very similar to VEs that we obtained directly from the dataset (Table 6).
Table 6: Comparing VE estimations from the selected BN model to the VEs estimated directly from the dataset. For **VE estimated from the BN model**, we estimated each VE 100 times using logic sampling inference algorithm and Equation 3. The average and standard deviation (SD) of 100 VEs used as the VE estimate and its uncertainty interval. For **VE estimated from data**, risk ratio test was used to obtain the VE estimates and 95% confidence intervals. P-values calculated by Fisher’s exact test.

<table>
<thead>
<tr>
<th>Condition</th>
<th>VE estimated from the BN model</th>
<th>VE estimated from data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VE</td>
<td>[VE−2×SD, VE+2×SD]</td>
</tr>
<tr>
<td>Overall</td>
<td>0.67</td>
<td>[0.59, 0.75]</td>
</tr>
<tr>
<td>Female</td>
<td>0.67</td>
<td>[0.57, 0.77]</td>
</tr>
<tr>
<td>Male</td>
<td>0.66</td>
<td>[0.52, 0.80]</td>
</tr>
<tr>
<td>AGE ≤ 45</td>
<td>0.72</td>
<td>[0.65, 0.80]</td>
</tr>
<tr>
<td>AGE &gt; 45</td>
<td>0.46</td>
<td>[0.17, 0.74]</td>
</tr>
</tbody>
</table>

Although we did not validate our results by reproducing them on other VE trial datasets, we expect that our proposed approach will perform similarly well; this is mainly because the machine learning algorithm could find the underlying statistical dependencies despite the little evidence in the data (only 1.6% influenza attack rate). We expect the proposed method to perform well on the datasets with higher attack rates. In addition to having a low attack rate, the dataset did not include children and very elderly subjects. This might be a reason for low Influenza incidents in the study as children and old population are the most vulnerable to influenza. The BN model of such datasets may not be usable in ABs of vaccination policy since these simulations usually include agents from all age ranges.

The significance of this research is that groups that conduct VE trials now have a method to release a statistical model of their data from which vaccination policy makers can obtain VEs conditioned on any set of individuals’ characteristics without the need to access the study dataset. Therefore, an AB of vaccination policy would be able to simulate the effects of a vaccine in presence of all characteristics of a population. We expect that an improvement in information about VE needed for vaccination policy analysis will lead to more effective use of vaccines, and ultimately improvements in health.

In the future, we will use an AB of vaccination policy to compare the new representation of VE trial datasets with the traditional method (using tables from publications). In particular, we will examine how the use of a statistical model of the VE trial dataset changes the results of an AB of vaccination policy that uses VEs from tables in the study publication.

**Acknowledgements**

We wish to thank the GlaxoSmithKline (GSK) company for providing the data used in this paper. This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award numbers R01GM101151 and U24GM110707 and by the National Library of Medicine of the National Institutes of Health under award number R01LM012095 and by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM111121. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the GSK company.

**References**


A Next-Generation Approach to Drug Diversion Monitoring

Jacob Smith, PharmD, MBA, Johns Hopkins Health System, Baltimore, MD, USA; Nick Culbertson, BS, Protenus, Baltimore, MD, USA

Abstract
Our country is in the midst of an opioid epidemic. Drug diversion plays a role in contributing to this as opioids are the most commonly diverted type of drug. In this session, presenters Nick Culbertson of Protenus, which has a data-driven platform that detects clinical drug diversion, and Jacob Smith of Johns Hopkins Medicine, which was a critical partner in building this one-of-a-kind platform, will help attendees understand why healthcare organizations cannot afford the reputational, monetary, or patient care costs associated with drug diversion. Additionally, they will discuss the evolution of clinical drug diversion programs and how proactive, AI-powered monitoring technologies maximize results and reduce effort when it comes to detecting and preventing this sort of unwanted activity. Attendees will walk away with an understanding of where they fall in the spectrum of approaches to drug diversion monitoring and short and long-term steps they can take to reduce this sort of inappropriate activity at their institution.

Presentation Outline

1. Intro - Overview of what we will be discussing
   a. The reputational, monetary, or patient care costs associated with drug diversion
   b. The evolution of clinical drug diversion programs
   c. How proactive, AI-powered monitoring technologies maximize results and reduce effort when it comes to detecting and preventing this sort of unwanted activity

2. Presenter Introductions
   a. Jacob Smith, Pharm. D.
      i. Johns Hopkins Medicine, Assistant Director, Medication Safety & Quality
   b. Nick Culbertson
      i. Protenus, Co-Founder & CEO

3. Overview of Drug Diversion Landscape
   a. What is drug diversion?
      i. The theft of prescription drugs from healthcare facilities for either 1) personal use/abuse or 2) re-sale on the black market. It’s a significant issue for hospitals, usually focused on front-line nursing staff (and anesthesiologists) who have the greatest direct access to drugs, and a significant contributor to the nation-wide opioid epidemic
   b. How common is it?
      i. Approximately 20% of the population ages 12 and older reports using prescription drugs for nonmedical use at least once in their lifetime.
Among health care providers, it is estimated that 15% of pharmacists, 10% of nurses, and 8% of physicians are challenged with alcohol and/or drug dependency.

c. What sorts of drugs are most commonly diverted?
   i. Commonly diverted drugs include opioids, stimulants, antiretroviral drugs, athletic performance-enhancing drugs, and non-opioid psychotropic drugs.
   ii. However, other types of pharmaceuticals are diverted as well that one may not expect.

d. How does drug diversion happen?
   i. Many health care employees have unrestricted access to controlled substances throughout the medication order filling and dispensing processes.
   ii. Hospital pharmacy leaders report increasing instances of diversion, including thefts of unopened vials, residual drugs left in a vial or syringe, and even substitution or dilution of medications intended for patients.
   iii. In one upsetting example, a cardiac technologist, who worked in 18 hospitals in seven states, diverted narcotics for self-use and infected an estimated 46 patients with hepatitis C.
   iv. 90% of reported theft or loss of controlled substances comes from pharmacies, and there have been over 12,000 thefts in the last three years.

4. Drug diversion has huge costs
   a. The estimated cost of controlled prescription drug diversion and abuse to both public and private medical insurers is approximately $72.5 billion a year.
   b. Economic cost related to drug abuse in the United States was estimated at $193 billion.
   c. In 1992, federal and state governments bore about $45.1 billion (46.2%) of the total $97.7 billion in drug abuse costs. Meanwhile, private insurance bore $3.1 billion, victims bore about $6.5 billion, and drug abusers and members of their households bore $42.9 billion.
   d. Approximately 100 individuals die from drug overdose daily, with opioids accounting for 75% of these overdoses.
   e. These incidents threaten patient and employee well-being, while also tarnishing the safety record of the hospital. Drug diversion risks patient harm if medications intended for patients are diluted or tampered. A health care provider diverting prescription drugs could even transmit pathogens, such as HIV or hepatitis C, through used syringes.

5. Where does monitoring fit into the whole drug diversion prevention strategies?
   a. Institutions approach this differently. Drug diversion will be held in different departments, but it’s most commonly under a drug diversion team, nursing or pharmacy.
b. Drug Diversion is heavily regulated. DEA, FDA, State DoH, State Licensing Boards, HHS OIG, State Attorney Generals all have some level of oversight, reporting requirements, or licensing oversight

6. Go through types of monitoring
   a. Complaint-driven
      i. An employee is reported for absenteeism and unusual behavior
      ii. Later learn that this employee pulled unusually high volumes of oxycontin and fentanyl
      iii. Conduct a reactive and time consuming manual investigation
   b. Report-driven
      i. Run a report to see employees who have pulled volume of narcotic three standard deviations above their peers
      ii. Hard to distinguish between real diversion activity and false positives
   c. Data-driven
      i. Ask and answer every question instantly
      ii. Utilizes artificial intelligence to look at the entire iceberg
      iii. Calculate levels of risk that each case poses to the institution

7. Evolution of Johns Hopkins Medicine’s Drug Diversion Program
   a. Discuss Drug Diversion team structure & core responsibilities
   b. Explain how monitoring system has changed over time
   c. Top reasons why they employ a data-driven monitoring approach, and how this has reformed their program to date

Educational Goals:
Attendees of this session will leave understanding:
- The most serious costs that clinical drug diversion poses to patients and their health outcomes, and healthcare organizations including hospitals and payers
- The evolution of drug diversion monitoring programs and their core differentiators
- Steps your organization can take now and down the road to begin taking a proactive approach to drug monitoring

Who should attend:

This session is designed to most directly inform stakeholders in drug diversion programs - diversion analysts and managers, nursing directors and managers, and pharmacists and pharmacy analysts. However, due to our data-centric approach that integrates artificial intelligence technology, this session will also be interesting and relevant to technologists, data scientists, biostatisticians, and healthcare innovators.
A Patient Centric Approach to Research Preferences in the EHR: An Innovative Partnership Between the Medical University of South Carolina, Yale School of Medicine, and the Epic Corporation

Leslie Lenert, M.D.1, Jihad Obeid, M.D.1, Melisa Habrat1, Allen Hsiao, M.D.,2 Tesheia Johnson2, Nancy Smider, Ph.D.2

1Medical University of South Carolina, Charleston, S.C., 2Yale School of Medicine, New Haven, CT, and 3Epic, Verona, WI

Abstract

Working to further a patient-centric model for research in the electronic health record (EHR), The Medical University of South Carolina (MUSC) and Yale have created unique and streamlined methods to capture patient preferences for participation in research through the Epic MyChart patient portal. Epic, after evaluation of many institutional workflows of their client base, has chosen to work with both MUSC and Yale to design and develop a combination of the two preferences designs into a future release for all Epic sites. Through this innovative partnership presentation, both MUSC and Yale will share the evolutions of their workflows and design and present data on effectiveness of their research preference implementations. Epic will share their rational for incorporating the two workflows into the standard functionality for Epic sites and what part it will play in research functionality within the Epic EHR starting in the 2018 version.

Partners-In-Innovation Panel Discussion

Introduction—Leslie Lenert (Moderator)

• Transition of patient research preferences from a static “demographic” item collected once-a-year, to a dynamic patient-controlled preference accessible at home and during clinical care.

MUSC and Research Participation Preferences—Jihad Obeid and Melissa Habrat

• Pre-EHR—Brief description of the paper-based research preference collection before the implementation of Epic with data from retrospective review data.
• First iteration—Brief description of initial implementation of electronic preference capture, data and decision behind redesign
• Development and implementation of phases of current contact and surplus specimen use model & implementation
  o Phase I (Patient Portal) collection of research preferences – implementation and continued maintenance with data
  o Phase II (In-Clinic) collection of research preferences– implementation and continued maintenance with data
• Contact management development and implementation
  o Workflow, rational, and benefits of having a research specific contact management workflow
• Lessened Learned, Ongoing Challenges and Future development

Yale’s Evolution of Research Preferences Capture—Allen Hsiao and Tesheia Johnson

• Pre-EHR
• Development of Research tab in Patient Portal
• “Help Us Discover” patient recruitment in Patient Portal, enabling patients to volunteer and save patient research preferences
• Journey towards Opt-Out Model that enables “Direct to Patient” Study Recruitment in Patient Portal
• Lessened Learned, Ongoing Challenges and Future development
Epic’s Integration of Research Participation Preferences into its Product—Nancy Smider:

- Basic design for research when first rolled out
- Current design and rational for research preferences collection
- Incorporation, design and development of the MUSC and Yale models into Epic for research preference collection

Conclusions—Leslie Lenert (Moderator)

- Importance of continued engagement with patients regarding research preference, and its effect on enrollment for clinical trials.
- Use of the preferences workflows to increase participant and effectiveness of both research and clinical patient reported outcomes (PROs).

Educational Goals:

- Understanding of how research preferences can increase patient engagement and recruitment of trials
- Understanding of the important difference between static and dynamic patient preferences for research
- Understanding of how innovative, creative development can influence highly utilized commercial vendors (i.e. Epic) can see the value for all client sites

Suggested Attendee Groups:

- Physicians and Investigators
- Directors, Managers and Analysts charged with research in the EHR
- Institutional leadership looking to implement or augment research in the EHR
Evaluating the utility of personal health assessment (PHA) data in predictive and outcomes research

Zachary B. Abrams, PhD; Hetian Bai; Kevin R. Coombes, PhD; Tasneem Motiwala, PhD, MPH

1Department of Biomedical Informatics, The Ohio State University

Introduction: Personal health assessments (PHA) are surveys consisting of questions related to an individual's health history, current health status, and lifestyle behaviors. Employers and health insurance companies commonly administer annual PHAs to understand the health risks of employees/covered entities and provide customized programs to reduce these risks. PHA questions are grouped under different categories that focus on distinct lifestyle behaviors e.g. physical activity, nutrition, smoking, alcohol, stress and well being; and health status e.g. chronic conditions. These questions are asked on a Likert scale of 1 (strongly agree) to 5 (strongly disagree), a scale commonly used in many social science questionnaires. Due to incentives offered to complete the PHA, they have a large response rate thus providing large amount of data with potential for use in predicting future health risks and healthcare expenses. However, to the best of our knowledge, PHA data have not been analytically tested to determine their value in predictive and outcomes research. In this abstract, we outline methodology for determining the experimental usefulness of PHA data, specifically focusing on how a cancer diagnosis can affect patterns of responses to PHA questions.

Methods: PHA data from 2013 and 2014 collected by The Ohio State University Health Plan was used in this study. A cohort of 24,987 members who completed PHAs in both 2013 in 2014 was separated into three experimental groups (cohorts): (1) members who do not have any cancer diagnosis (N=24,443), (2) members who had a diagnosis of cancer following completion of their 2014 PHA (N=279), and (3) members who had a diagnosis of cancer in-between filling out their 2013 and 2014 PHAs (N=265). By observing how these three groups answered their PHA questions we can assess whether having a severe disease such as cancer changes how an individual answers their PHA.

Prior to conducting the analysis, we performed an evaluation of the structure of PHA responses. Overall the PHA has a limited ontology in terms of categories as discussed in the introduction. We observed a high level of correlation in responses both across categories and particularly within categories (data not shown). While this may be intentional to create redundancy across questions for internal validation, this redundancy presents a problem in data analysis, causing inflation of the signal while suppressing variability. We, therefore, decided to focus on a limited set of non-correlated questions measuring well-being since it is a measure that is associated with several health-related outcomes1. Based on the multiple dimensions of well-being2, we focused on three question categories for feature reduction: (i) emotional health status, (ii) physical health status, and (iii) work-life balance. Within each of the three categories, the three questions with the smallest combined Pearson correlation scores were selected.

This was done to reduce the redundancy and increase the variability across the three questions selected in each category. The composite score for each category for each individual member was calculated as the sum of the responses to the three questions in the category (potentially ranging from 3 to 15) for their 2014 PHA, with higher scores indicative of poorer well-being. These scores indicate the level of agreement or disagreement an individual has for the question...
group. The scores were compared across the three cohorts utilizing a student’s t-test (Table 1). Only one PHA question group, physical health status, had significant differences between cohorts (P=0.02 and P=0.03 for cohort comparisons 1-2 and 1-3, respectively). This indicates that individuals that have been diagnosed with cancer or are developing cancer differentially have a lower outlook on their own health than the rest of the population.

**Table 1: Comparison of scores between the three cohorts**

<table>
<thead>
<tr>
<th>Question Group</th>
<th>Mean Scores</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohort 1</td>
<td>Cohort 2</td>
</tr>
<tr>
<td>Emotional Health Status</td>
<td>4.87</td>
<td>4.88</td>
</tr>
<tr>
<td>Physical Health Status</td>
<td>6.11</td>
<td>6.49</td>
</tr>
<tr>
<td>Work-Life Balance</td>
<td>5.77</td>
<td>5.77</td>
</tr>
</tbody>
</table>

We also assessed the variability of responses to this set of PHA questions across years. To quantify an individual’s variability from year-to-year, we calculated the absolute difference between an individual's answers for each question from 2013 to 2014. Since every PHA question is on a 1 to 5 scale, a value between zero and four was possible. Adding all absolute difference values for an individual creates a single score quantifying the amount of change in how a person answered the PHA from 2013 to 2014. These scores were compared between each of the three cohorts using a student’s t-test. While cohorts 2 and 3 have a lower variability than cohort 1, the only significant difference was between cohorts 1 and 2 (P=0.03 and P=0.13 for cohort comparisons 1-2 and 1-3, respectively).

**Results and Discussion:** Our analyses indicate that cancer patients consistently rate their own health as lower than the rest of the population. The lower variability year to year is of particular interest since individuals in cohort 2 answered their 2013 PHA without them knowing they had cancer. This means they were consistently indicating poor health prior to the cancer diagnosis. This similar effect appears to be occurring in cohort 3 since they also have much lower, though non-significantly different, variability than the rest of the population.

**Conclusion:** Personal health assessment (PHA) data can be useful as a form of self-reported medical data. Our study found that members with cancer had overall higher scores for physical health-related questions indicating poorer assessment of their own health than the rest of the population. They also had lower variability in answering PHA questions between two years demonstrating that they tended to consistently indicate poorer health than did the rest of the population. In the future, we hope to follow-up this research by looking at other chronic diseases such as cardiovascular disease to test whether similar trends can be seen with other diseases.

**References**

Introduction

Blockchains are distributed ledgers underlying Bitcoin and the subsequent cryptocurrencies it spawned. Data in blockchains are bundled into blocks, with each block containing a hash of its predecessor. This hash chains the two blocks together. The resulting chain of blocks is resistant to tampering, as changing the data in one block would invalidate that block and all the subsequent ones. Also the distributed nature of blockchains endows them with the traditional strengths of distributed systems such as fault tolerance and consequent robustness. As employed in Bitcoin, blockchains also achieve disintermediation by eliminating the need for a centralized authority.

The application of blockchain to health care is an extremely active area of research. Solutions that use blockchains in Health Information Exchanges and distributed predictive modeling have been proposed. Benchoufi et al conclude that blockchain may be the key to addressing the challenges of contemporary clinical research, and “should draw the attention of the whole clinical research community.” In this paper, we propose a conceptual model for the application of blockchains to clinical research, and apply this conceptual model to the management of the Electronic Trial Master File (eTMF) in a multicenter clinical trial.

Methods

Conceptual Model for Blockchain Innovations in Clinical Research

Our conceptual model (Figure 1), has four elements/layers: nodes, transactions, data, and logic. This model can be used to map entities in the clinical research ecosystem – such as stakeholders, systems and processes – to blockchain elements for the effective design of solutions.

Nodes: These are the participants in the blockchain implementation. In clinical research, nodes represent stakeholders such as patients or researchers; systems such as electronic health records (EHRs) or integrated data repositories (IDRs); and entities such as sponsors or contract research organizations (CROs).

Transactions: Clinical processes or interactions between two nodes are considered as transactions. One example might be a query sent to an IDR. A transaction might involve exchange of data originating from a clinical research process, or it might involve exchange of metadata about a process. Transactions have a clear beginning and end.
**Data:** This layer encapsulates the data (or data about the data) generated, managed and exchanged in a blockchain. This could represent data in EHRs, clinical trial management systems (CTMS) or an IDR. Data can be distinguished from transactions by the fact that data are permanent. In clinical research applications with high emphasis on audit trails, every transaction – whether successful or not – will become data.

**Logic:** This layer, similar to Smart Contracts, represents business logic that translates to actions. These are triggered when a certain kind of transaction is activated between two nodes. For example, documenting an adverse event could trigger logic that notifies the Principal Investigator (PI) and Data Safety and Monitoring Board (DSMB). Logic could also represent what a node is permitted to do and how it interacts with other nodes.

**Results**

**eTMF Management in a Multi-site Clinical Trial**

The Trial Master File (TMF) for a clinical trial is a collection of essential documents that the sponsor or coordinating center (such as the DCRI or a CRO) and investigators/sites have to maintain under the International Conference on Harmonization (ICH) Good Clinical Practice (GCP) guidelines. Essential documents that are part of the TMF can be inspected by a monitor or regulatory authorities to validate the integrity of the clinical trial data, making their timely maintenance important.

In the case of maintaining an eTMF on a blockchain in a multicenter clinical trial, the coordinating center, sponsor or CRO is a node. Each of the study sites and the associated site IRBs are nodes, along with the DSMB and FDA. Transactions consist of the addition of each essential document and its subsequent versions to the eTMF, and any requests from other nodes to review or verify any part of the TMF. The essential documents in the eTMF, along with identities of signatories and associated audit trails, comprise the data layer. Finally, the logic layer represents how each of the nodes can only add or read certain essential documents (Figure 2). For example only the sponsor can add the clinical study report document to the TMF, while the FDA can only read it.

Maintaining eTMF on blockchain could potentially reduce monitoring costs if documents such as signed informed consents could be verified via blockchain while preserving privacy. This could be valuable as the cost of monitoring can account for as much as 14% of the total cost of a trial. eTMF on blockchain can eliminate fraud as the data is immutable, thereby securing data provenance. This can also speed up the process of preparing submissions to FDA.

**Discussion**

The proposed model can help design applications that leverage blockchain’s strengths to address acknowledged issues in clinical research, and specifically in clinical trials, such as the high cost of monitoring, contracting and source document validation, and trust issues endemic to globalized research. There is also a need for an evaluation model to help researchers weed out blockchain applications driven by hype, and understand the implementation challenges.

**References**


Imon Banerjee, Ph.D.1, Michael Francis Gensheimer, MD2, Douglas J. Wood, MS1, Solomon Henry, MS1, Daniel Chang, MD2, Daniel L. Rubin, MD1

1Department of Biomedical Data Science, Stanford University, Stanford, CA,
2Department of Radiation Oncology, Stanford University, Stanford, CA

Introduction

Scientific review(1) and clinical studies(2) have indicated that physicians are overly-optimistic during survival prediction of patients with terminal cancer. This leads to overutilization of aggressive medical interventions and protracted radiation treatment, which increase side effects and health care bills, while other patients who may benefit for continued therapy are likely under-treated. The electronic medical records systems serving major medical centers store a wealth of potentially valuable clinical notes acquired over time (“longitudinal data”) that contains rich information on performance status, imaging findings, and tolerance to systemic therapy. These data contain key covariates for informing diagnostic decisions, and may also serve as critical resources for survival estimation. However, the massive explosion of the medical data outstrips the manual ability to comprehend the entire source of information. On the contrary, a machine learning model can be trained on the large amount of heterogenous data which allows seamless integration between multi-source information, and, as a result, it may outperform physicians in estimating life expectancy. Using the large Stanford Cancer Institute Research Database (SCIRDB), we created a dynamic sequence-dependent deep learning model - Probabilistic Prognostic Estimates of Survival in Metastatic Cancer Patients (PPES-Met). The PPES-Met model takes as input a sequence of longitudinal free-text clinical visit narratives ordered according to the date of visits, and computes as output a probability of short-term life expectancy (> 3 months) for each visit. The complexity of the model mainly lies in extracting relevant information from the heterogeneous types of free-text clinical notes along with modeling temporal irregularity of the visits.

Methods

Patient population: With the approval of Institutional Review Board (IRB), we created a database that includes adult cancer patients (13,523) seen at the Stanford Cancer Center from 2008-2017 and diagnosed with distant metastases. This database contains various types of free-text patient visit notes (e.g. radiology reports, oncologist notes, discharge summary) from date of metastatic cancer diagnosis to death. A separate “Palliative radiation dataset” was created using patients (899) enrolled from 2015-2016 in a prospective survey study conducted in our institution’s Radiation Oncology department. The overall group of patients were seen for 471,005 daily encounters/visits, including outpatient and inpatient contact. For these visits, median follow-up was 12.7 months. Median overall survival was 22.4 months. Patients were hospitalized for 115,716 (24.6%) visits. There were 1,403,544 provider notes. The training set contains 10,239 patients with 380,080 visits, validation set of 1,785 patients and test set of 1,818 patients (15%) with 90,925 visits.

Proposed System - PPES-Met: The model is composed of two core processing blocks:

i. Semantic Word Embedding (SWE): We adopted a completely unsupervised hybrid method – an updated version of Intelligent Word Embedding (IWE) method(3) that combines semantic-dictionary mapping, neural embedding, and context-based windowing technique for creating dense vector representation of free-text clinical narratives. The method leverages the benefits of unsupervised learning along with expert-knowledge to tackle the major challenges of information extraction of informative information from clinical texts, while accounting for ambiguity of free text narrative style, lexical variations, arbitrary ordering of words, and frequent appearance of abbreviations and acronyms.

ii. Stacked RNN model: On top of the word embeddings obtained from patient visit notes, we designed a many-to-many RNN model using two-layer one directional stacked stateful Long short-term memory (LSTM) units for learning survival across the sequence of clinical narratives. The model takes as input a series of vectorized patient visit notes ordered according to the timestamp of visits, and it predicts probability of survival at each patient visit. In the stacked RNN layers, the first layer’s one directional LSTM block receives the input $x_t$ and previous hidden state $h_{t-1}$, and pass the current state $h_t$ to the successive LSTM block and to the corresponding block in the upper layer. LSTM blocks also maintains an H-dimensional cell state $C_{t-1} \in \mathbb{R}^H$ and the second layer units have modeled to maintain the recurrent connections in multiple dimensions. The final output of double layer stacked-RNN is modeled as: $\hat{y}_t^{(1)} = f_h(h_t, \hat{y}_{t-1}^{(1)})$ where $h_t$ is the hidden state of the first level and $\hat{y}_t^{(1)}$ is the predicted survival at time $t-1$. The time distributed weighted cross entropy loss is used as optimization function during model training. By using stateful LSTM, the model includes long-term dependencies that exist in the longitudinal data, which is generally very informative for the prediction task.
Results

Quantitative evaluation: We used 3-month survival record defined at each time point as categorical class labels for probabilistic prediction. The overall prognosis estimation accuracy on the test dataset was validated as Precision-Recall curve and AUC score was 0.97. The high AUC scores show that the model is predicting survival accurately with high precision, as well as returning most of positive results (high recall). To check the calibration with the ground truth, we also measure the Brier score and the low brier-score (0.069) shows the prediction was highly calibrated with the ground truth.

Figure 1 Intelligible longitudinal survival curve of a patient

Qualitative evaluation: To provide a formalized mechanism to reason about computerized model’s predictions at a specific timepoint, we implemented an interactive graphical interface that generates a longitudinal probabilistic summary (Figure 1) for each patient. Clicking on a time point, the system will retrieve not only the visit type, but it will also exploit the controlled-terms and will extract the core findings of the visit by highlighting the context of the controlled-terms (see Figure 1). This intuitive illustration may help the clinician to reason on the PPES-Met prediction and perform a qualitative error-analysis.

Discussion

The objective of our work is to improve physicians’ knowledge of their patients’ prognosis to help tailor treatment strategy, improve quality of life, as well as reduce costs. Early stage results of an ongoing prospective study conducted by Stanford Radiation Oncology department which includes 899 patients enrolled in the palliative radiation study found that the physicians are not able to accurately estimate life expectancy. We tested our PPES-Met model on a combination of a general group of metastatic patients and data from a palliative radiation study, and the probabilistic prediction accuracy was 0.97 AUC-PR-curve. Initial evaluation suggests that the PPES-Met prediction model produces good accuracy. This is probably due to the PPES-Met model’s capability of integrating a large amount of patient-specific facts while preserving long-term dependencies in the data, which is not a trivial task for human experts. The high accuracy and the ability of our visualization the PPES-Met model output to show the critical data driving its predictions suggests that the model might ultimately be clinically useful in the future as a decision support tool to personalize metastatic cancer treatment and to aid physicians. The core limitations of the current work are: (i) many patients in the training dataset lost follow-up which may create an inaccurate assumption about survival during the model training. However, we consulted the central cancer registry to validate the survival; (ii) the time points are not equally spaced, and a single day may contribute multiple data points which may affect the temporal dependency and introduce fluctuation in the survival curve. In future, an NLP technique can be applied to combine the visit data from the same day for date-based survival analysis.

References

Improving patient and caregiver engagement through the application of data science methods to audio recorded clinic visits stored in personal health libraries using ORALS.

Paul J Barr, MSc, PhD,1 Saeed Hassanpour, MMath, PhD,2,3,4 William Haslett, PhD,2 Michelle D Dannenberg, MPH,1 Craig H Ganoe, MS,2 Rebecca Faill,2 Amar Das, MD, PhD,1,2 Roger Arend, BS,5 Meredith C Masel, MSW, PhD,6 Sheryl Piper,7 James Ryan, DO,9 Martha L. Bruce, MPH, PhD,1,10 and Glyn Elwyn, MSc, MD, PhD1

1 The Dartmouth Institute for Health Policy and Clinical Practice, Geisel School of Medicine, Dartmouth College, Lebanon, NH. 2 Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH. 3 Department of Computer Science, Dartmouth College, Hanover, NH. 4 Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, NH. 5 Patient Partner, Hanover, NH. 6 Oliver Center for Patient Safety and Quality Healthcare, University of Texas Medical Branch, Galveston, TX. 7 Patient Partner, Ludington, MI. 9 Ryan Family Practice, Ludington MI. 10 Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Lebanon, NH.

Introduction
Forty to eighty percent of healthcare information is forgotten immediately by patients after their visit.1 This is a significant barrier to self-management, a central component of the Chronic Care Model, resulting in poorer health outcomes.2 An after-visit summary is a common strategy to improve recall-of-visit information, but imposes a significant burden on clinicians who must document the entire visit in terms that are understandable to patients.3 Alternatively, audio recordings can provide a full account of the clinic visit, improving patient understanding and recall of visit information.4

As patient demand for recordings increases, a growing number of clinics across the U.S. are offering audio recordings of clinic visits.5 Yet, unstructured clinic recordings may overwhelm patients. To address this challenge, we are developing the Open Recording Automated Logging System (ORALS), a platform that will provide structure to clinic audio-recordings.

Beyond the typical recording and playback functions of existing recording software, ORALS will offer the following features (Figure 1): 1) automatically transcribe clinical encounter audio recordings; 2) use of Natural Language Processing (NLP) techniques to identify key moments in the recording (e.g. discussion of medications). In the ORALS playback interface, these moments will be tagged in order to support efficient navigation of the recording; 3) use of NLP techniques to associate important content from the visit to credible resources on the Internet, such as Medline Plus; and 4) enable the sharing of recordings with third parties, such as caregivers.

An illustration of ORALS is shown in Figure 1. Consumers will record their clinical encounters using a mobile device, and will interact with those recordings using either a mobile device or a computer. The system will be built as a web application, with all data stored on the web server (recordings will be deleted from user devices once they are uploaded). On the server side, the application will transcribe incoming recordings, tag important moments in the conversation for the user, and will gather relevant information from trusted sites. Finally, the processed recordings, with their tags and resource links, will be presented to users in a simple playback interface.

Methods
Adopting a user-centered design framework, we are engaging end users throughout the development of ORALS. We began with a case study of three practices that routinely share clinical encounter recordings. We have conducted 68 interviews including 31 patients, 8 caregivers, 16 clinicians and 3 programmers in order to understand implementation strategies, barriers and facilitators of recording, and identification of key visit information.
For the user interface (UI), we are conducting focus groups and one-on-one usability with patients and caregivers. While this is occurring, development of the back-end system is underway. The ORALS UI will be developed iteratively, with results of end-user usability work informing the design of each prototype cycle. Formal usability evaluation will be performed in a human-computer interaction lab to ensure that ORALS surpasses acceptable usability metrics guided by the TURF (Task, User, Representation, and Function) framework.5

Our text-mining and machine-learning (ML) methods will include regular expressions, vector space modeling, distributional semantics neural networks, and logistic regression. Our sample size for developing the ML models will be calculated through statistical power analyses to ensure it is sufficient to train reliable and accurate models.

ORALS will be evaluated in a randomized, controlled pilot trial. Participants (N=70) will receive either a health library with ORALS or a generic personal health library with no recording. We will measure usability, satisfaction, patient activation, and caregiver confidence at one month. If the results of this pilot study are promising, we will pursue a larger randomized trial, aiming at improving self-management capabilities and outcomes for patients.

Our NLP development begins with the transcription and annotation of approximately 200 clinical encounter recordings that we have on-hand, with a minimum of 200 additional recordings to be collected over the next 2 years. These annotated transcripts will provide the initial basis for an NLP model that extracts clinically important information and recommendations. Although this type of extraction has been done in the past,7 previous efforts are limited by their dictionary-based and rule-based information extraction methods, reducing their generalizability in broad medical domains due to natural language variations and ambiguity. We plan to leverage novel distributional semantic features in combination with various syntactic features and powerful ML frameworks to develop new medical text analysis methods to address the shortcomings of current text analysis methods.

Results
Preliminary findings from our case study reveal that access to recordings “tagged” for key information would be desirable, including symptoms/diagnosis, medication and treatment plan. Preliminary usability work with paper prototypes has found that for information-finding tasks most participants will use the tags to invoke audio playback over other standard audio playback interactions (e.g. play button) and has aided in new feature suggestions, such as highlighting the tag(s) of the audio segment currently being played. We have built a proof-of-concept web application that supports secure audio recording and playback of clinical encounters. Annotation of the aforementioned recordings is also underway, and the data pipeline for our NLP techniques is being developed.

Discussion
The proposed ORALS platform will address poor recall and challenges in seeking health care information for patients and caregivers, which contribute to poor health outcomes. The development of an open access, innovative personal health library, with the novel application of NLP to create easy-to-navigate recordings of clinic visits connected to trusted patient resources aligns with patient demand and can improve the quality of patient health information seeking, improving patient self-management capabilities.

References
Adding data visualizations of healthcare system dynamics to i2b2

Nick Benik¹, Daniel Bolin², Katy Borner, PhD², Nick Brown¹, Scott Hutcheson², Griffin Weber, MD, PhD¹

¹Harvard Medical School, Boston, MA; ²Indiana University, Bloomington, IN

Introduction

Electronic health records (EHR) and administrative claims databases are transforming medical research by giving investigators access to data on millions of individual patients¹. Compared to manual paper chart review, these databases reduce the time and cost of clinical studies by orders of magnitude, enabling types of research that were unfeasible in the past. However, investigators often incorrectly treat EHR and claims data as simply big versions of clinical trials data. Yet, there are important differences: During clinical trials, patient information is obtained and recorded in a standardized way and checked for accuracy and completeness. In contrast, EHR and claims are observational databases, which reflect not only the health of the patients, but also their interactions with the healthcare system. For example, the date associated with a code for diabetes is when the physician made the diagnosis, not when the patient first developed the disease. These observations are influenced by the dynamics of the healthcare system—when physicians schedule visits with their patients, which tests physicians decide to order, what codes need to be recorded to get reimbursed for procedures, etc. By ignoring this dimension of the data or naively treating it as noise, investigators risk both misinterpreting the actual patient pathophysiology and losing valuable information content.

Last year we gave a presentation at AMIA showing how information about patient health can be obtained through large-scale analysis and visualization of “healthcare system dynamics” (HSD). A simple example illustrates how it works. The value of a white blood cell (WBC) count laboratory test result is an observation of the patient’s pathophysiology. The time of day when the test is ordered is an observation of HSD—a physician decided a WBC count was needed. It is counterintuitive, but we have found that patients with an abnormal WBC value at 4pm are actually more likely to be alive after 3 years than patients with a normal WBC value at 4am. This may simply be due to physicians ordering routine blood counts for healthy patients in the afternoon, but only ordering counts at 4am when they believe patients are sick.

This year we describe our work creating software extensions to the i2b2 (Informatics for Integrating Biology and the Bedside, www.i2b2.org) platform to help investigators incorporate HSD into their research. i2b2 is a widely adopted open source software platforms for querying clinical data repositories. We added concepts to the i2b2 ontology related to HSD, and we developed several data visualization “plugins” for i2b2. Our presentation will introduce these extensions to the audience and demonstrate their functionality. We will also discuss the technical challenges and design choices we made when implementing these extensions. Specifically, i2b2 has many ways of defining ontologies and building new plugins. They have tradeoffs on usability, performance, storage requirements, and software complexity. The approach we used could help others determine the best way to build their i2b2 extensions. (Dozens of i2b2 plugins have been developed by various groups around the world.)

Methods

i2b2 is an ontology based system, which includes several standard medical vocabularies, such as diagnoses (ICD-9, ICD-10), laboratory test results (LOINC), and medications (RxNORM). Note that these only enable users to query the pathophysiology dimension of the data. We identified an initial set of HSD concepts/variables and developed an i2b2 ontology for these items. We implemented each item in the ontology in i2b2 in several different ways to understand their benefits and limitations. We also created data visualization plugins for i2b2 to help users understand the concept of HSD and see its influence on their datasets. We implemented these in two ways: first as a client-side plugin that performs all the data processing in the user’s web browser, and then as a both a client-side plugin and a server-side plugin so that some of the data processing can occur in the database. The software is available at http://HealthcareSystemDynamics.org.

Results

HSD Ontology
We added concepts related to time periods to the HSD ontology. Time periods include the month of the year (e.g., January, February, etc.), the day of the week (e.g., Sunday, Monday, etc.), or the hour of the day in which an observation occurred. In the past, i2b2 could only query by date range. For example, users could select patients who had influenza between Oct 1, 2015, and Oct 31, 2015. However, there was no easy way select patients with influenza in October of any year. Defining time periods in the ontology as functions applied to the existing date fields in i2b2 did not require any database schema changes, but queries run slowly since the time periods must be calculated in real time. Creating a new time dimension to the database greatly improves performance, but extra work is needed to build and update an additional database table.

We also added “fact count” to the ontology, which represents the total amount of data that a patient has (i.e. all the diagnoses, medications, lab values, etc.)\(^2\). We have previously applied fact count to a variety of use cases, such as predicting survival, identifying control cohorts, and estimating data completeness. However, i2b2 has never had a built-in way of determining fact count or using it in queries. We added fact count in two ways. First, we treated fact count like a laboratory test, where a user is prompted to type in an arbitrary numeric value range to query. Then, we created separate ontology concepts for pre-defined percentile ranges. The lab test approach is more flexible, but the pre-defined ranges can take advantage of i2b2’s “breakdown” functionality (e.g., breakdowns by age, race, etc.).

**HSD Visualization**

We created two sets of data visualization plugins for i2b2: A laboratory test heat map (Figure 1a) shows how a user-selected outcome (e.g., color = survival) varies by both the value of the test result (columns) and different time periods (rows = hour, weekday, or month). A fact count visualization (Figure 1b) shows the distribution of patient fact counts by age, with overlays of different cohorts of patients. We leveraged two capabilities of i2b2 in building these visualizations: client-side plugins for the i2b2 web client, and server-side “cell” plugins. We found that server-side plugins are more difficult to test and deploy in a production environment, especially in a tightly controlled hospital setting. However, they are far more scalable, enabling functionality that is not feasible with only i2b client-side plugins.

**Figure 1: Time of day and fact count visualizations.**

**Conclusion**

In observational databases, each data point is a combination of both patient pathophysiology as well as HSD, and both of those dimensions provide valuable information for research. We showed how we extended the i2b2 platform to help investigators incorporate HSD into their research.

*This work was funded by NIH/NCI grant U01CA198934.*

**References**

Neighborhood Socioeconomic Status as a Predictor of Health Outcomes in Electronic Health Record Based Studies

Nrupen A. Bhavsar, PhD1, Aijing Gao2, Neha Pagidipati, MD1,3, Matthew Phelan, MS3, Benjamin A. Goldstein, PhD2,3
(1) Department of Medicine, Duke University School of Medicine, Durham, NC
(2) Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC
(3) Center for Predictive Medicine, Duke Clinical Research Institute, Durham, NC

Introduction: There is growing interest in the link between the social and economic context within which a person lives and the impact that this may have on their health. Recent studies have reported an association between neighborhood residence and multiple health measures, including chronic disease, infectious disease, and more generally, healthcare utilization.1-3 There is less information on the relative utility of neighborhood socioeconomic status (nSES) in models that predict future health outcomes. To that end, we used socioeconomic data that is linked to the electronic health records of patients seen within the Duke University Health System (DUHS) to quantify if: 1) nSES, as defined by census tract-level SES, is associated with poor health outcomes; 2) data elements in the EHR are correlated with census tract-level SES data (i.e., can be used for phenotyping); and 3) census tract-level SES data alone or in concert with EHR data can improve risk prediction beyond current models by using EHR data.

Methods: Data Source: We used data from the DUHS EHR system from 2009-2015. The DUHS consists of two community hospitals, one large referral hospital, and a network of outpatient clinics. The patient population included 187,859 patients split temporally into a training (n=81156) set that includes patients from 2009 (with outcomes from 2009-2012) and a validation set (n=106703) that includes patients from 2012 (with outcomes from 2012-2015).

Neighborhood Socioeconomic Status: nSES was derived using census tract level data collected by the American Community Survey (ACS). The 5-year ACS estimates of 1) the percent of households that average ≥ 1 persons/room, 2) median value of owner-occupied dwelling, percent unemployed, 3) percent living below poverty level, 4) median household income, 5) percent ≥ 25 years of age with bachelor’s Degree or more, and 6) percent ≥ 25 years of age with < 12th grade education, were used to calculate the Agency for Healthcare Research and Quality (AHRQ) SES index.

Demographic and Clinical Measures: Information on demographic (e.g., age, race, sex), clinical (comorbidities), laboratory (e.g., cholesterol), medication (e.g., statins), vitals (e.g., blood pressure) and service utilization (e.g., emergency department encounters) was obtained through the Duke University Medical Center EHR. Outcomes of interest included accidents, asthma, influenza, myocardial infarction (MI), stroke, venereal disease, and service utilization (emergency department, inpatient, and outpatient encounters). These outcomes were assessed over 6 time horizons: 30 days, 90 days, 180 days, 1 year, 2 years, and 3 years. For the purposes of this abstract, only results for select outcomes are presented.

Statistical Analysis: The AHRQ SES score was modeled linearly and subsequently stratified using pre-defined quartiles. Survival analysis was used to quantify the association between nSES and risk for asthma, influenza, myocardial infarction, and emergency department encounters. Logistic regression models were fit for each outcome over 6 time horizons using least absolute shrinkage and selection operator (LASSO) for model selection to determine if census tract data improved discrimination. EHR data was fit to the predictive models. Constructs defined using nSES quartile were then added to the prediction model. C-statistics was used to assess discriminatory ability of the predictive models. All statistical analyses were performed in R 3.1.4.

Results: Patients in lower nSES quartiles had higher risk for MI and Stroke: The probability of MI and stroke were greater for patients living in lower quartiles of nSES as compared to higher quartiles of nSES (Figures 1a and 1b).
nSES is correlated with EHR variables, mainly due to demographic variables: Variables in the EHR or demographic variables only were regressed on nSES. The majority of the correlation between EHR variables and nSES ($r^2=0.31$) was explained by demographic information ($r^2=0.29$) within the EHR ($p<0.01$). In LASSO models that incorporate EHR and SES data at a time horizon of 3 years, EHR variables such as demographics, service utilization, history of medication use, and comorbidities were frequently selected while variables such as the AHRQ SES index were rarely chosen. The relative importance of nSES varies by clinical outcome of interest, age, and time horizon.

Time horizons: To determine the added value of census tract data, we used nSES alone or in concert with EHR data to fit the models over various time horizons. For outpatient encounters we did find prediction improved as the time horizon increased (Figure 2a) while for ED encounters, prediction decreased as the time horizon increased (Figure 2b); however, for most other outcomes, we found no significant enhancement in c-index in models with EHR alone as compared to models with EHR and nSES data combined.

Age: The risk models for predicting ED encounters in middle-age adults had enhanced discriminatory power after adding nSES data (Figures 2b). Similar improvements in discriminatory power were not seen for other outcomes.

Discussion
Our study shows that the added value of nSES is less than we would expect as much of the variability in nSES may be phenotyped through demographic information recorded in the EHR. In discrete instances though, nSES can improve risk prediction. In our presentation, we will highlight the data sources and methodology used to come to this conclusion and the discrete instances where nSES may help in risk prediction.

References
Patient Opinions About Digital Messaging for Clinical Research Recruitment

Michael N. Cantor1,2, MD MA, Michael Mishik2, B.E.,
1NYU School of Medicine, New York, NY USA; 2Medical Center Information Technology, NYU Langone Health, New York, NY USA

Abstract
Little published research exists about patients’ attitudes toward the use of patient portals as tools for clinical research. We used our patient portal to survey over 2100 active patients about their attitudes toward electronic, research-related communications. 87% of respondents said that they would be interested in receiving research recruiting messages through the portal, and approximately 45% of patients said that there should be no limits on the number of recruitment-related messages they receive. EHRs and patient portals are valuable tools that should be used to speed enrollment in clinical trials, and a large group of patients would like to see information on all applicable trials, without limits.

Introduction
Identifying and recruiting patients for clinical research remain two of the biggest challenges for researchers. EHRs are playing larger roles in recruitment, from helping identify cohorts to alerting clinicians about potentially eligible patients either during or before scheduled visits1,2. The ubiquity of technology in patients’ lives has led to the incorporation of electronic communications, from email to Facebook ads, as important complements, or even substitutes, for traditional recruitment techniques (flyers, mass mailings)3,4. Several academic medical centers have successfully expanded the use of their EHR systems beyond the usual clinical messaging to incorporate both communication and registry functions related to research5,6.

One major issue around using EHRs for recruitment that remains unsettled is the best way to manage the amount and type of messaging patients receive. For example, if a patient opts-in to a research registry, should they receive alerts about every study for which they are potentially eligible? If not, how should one limit the number of alerts and, how does the institution balance patient autonomy with an approach that could be seen as paternalistic? Few institutions have set policies around the number of alerts a patient should receive. Before setting policy at NYU Langone Health, we chose to survey our patient population to better gauge their beliefs and to obtain data that would help us formulate a policy that aligns with patients’ actual preferences rather than their hypothetical concerns.

Methods
The NYU School of Medicine IRB determined that this project was exempt from review. We developed a 10-question survey (6 questions related to digital communications [Table 1], 4 related to demographics) in our institution’s RedCap system. The survey was developed with input from our institutional Research Governance Group, who gave input on both content and wording of the questions and was approved by NYULH’s institutional communications group. We then obtained a random sample of 20,000 adult, active MyChart (our Epic® patient portal) users, and sent a survey invitation via MyChart’s secure platform to 10,000 patients per week for two weeks. The survey was completely anonymous, and the invitation contained a direct link to the RedCap survey.

Results
Demographics: We received 2157 responses to the survey, 2154 of which were complete. 61.7% of respondents were female, 83% were white, and 11% identified as Hispanic/Latino. 2/3 of patients were within the 46-75 age range.

Content: 72% of patients responded that they would be interested in participating in research studies, and 87% of all patients responded that they would be interested in receiving research-related messages, for studies where they would be a potential candidate, through MyChart. Responses about limits on the number of messages that patients receive were nearly evenly split, with 46% responding that there should not be a limit on the number of research-related messages received, and 39% responding that there should be a limit. Over 90% wanted the option to opt-out from receiving further messages with each invitation and were willing to give the opt-out reason. Opinion about the “right” number of research-related messages that a patient could receive in a particular time period was mixed (“If you received messages in MyChart, do you think there should be a limit on the number of messages you receive related to research during a specific time period…), as seen in Figure 1.
Discussion
Because we surveyed patients who were already signed up for our clinical portal, and the population of respondents was not very diverse, our results may not be generalizable to the population who may be eligible for research studies. Though we could have kept the study open longer and contacted many more patients, even with our low response rate, we expected that we would obtain a statistically valid sample at this number of patients. Institutions should determine the number of patients necessary to obtain a large enough sample on which to base policy decisions. The significant number of patients who wanted no limits on research-related communications points to the importance of patient autonomy and its balance with institutional priorities. Creating a flexible system that allows patients both to opt-out and to control their exposure to communications would provide an optimal solution. Future work will evaluate the effectiveness of digital messaging in terms of actual enrollment.

Conclusions
Developing and distributing the survey through the patient portal was a relatively straightforward process and was relatively quick to implement after institutional approval. Our survey revealed engaged patients who are interested in finding out more about research, and who wish to have input into and control over the number and types of research-related messages they receive. Performing a similar survey in a larger, more diverse population would help give a more accurate picture of broader patient preferences. Digital messaging can be an effective tool for research-related communications, and could be added to the current toolbox of general recruitment techniques.

References
Open mHealth: Common Data Schemas and API for Mobile Health Data

Simona Carini, MA1,2, Emerson Farrugia, MS2,3, David Haddad, MSc2,3, Deborah Estrin, PhD4,2, Santosh Kumar, PhD5, Ida Sim, MD, PhD1,2
1University of California San Francisco, San Francisco, CA; 2Open mHealth, San Francisco, CA; 3Overlap Health, Inc, Oakland, CA; 4Cornell Tech, New York, NY; 5University of Memphis, Memphis, TN

Introduction

Mobile phones and devices, with their constant presence, data connectivity and multiple embedded sensors, are producing tremendous amounts of data that can provide fine-grained, moment-by-moment views of physiologic, environmental, and behavioral features of health and disease states. Such personal mobile health (mHealth) data streams are valuable for understanding complex multifactorial diseases, especially when they can be combined with electronic health record (EHR), genomic and other data for discovery, health improvement and disease management. Hence, mHealth data interoperability through data and metadata standards is a necessity.

Reminiscent of the siloed electronic health record (EHR) market, mobile health has been developing as a collection of incompatible systems in which data are read and written in inconsistent formats that are hard to integrate and use in meaningful ways to improve health, manage disease and make biomedical discoveries.

Compared with data integration for purely consumer-facing solutions (e.g., fitness and lifestyle apps), integrating mHealth data into clinical care requires additional attention to semantic clarity. For example, for a blood glucose value to be clinically meaningful, it should be temporally contextualized with food intake. Open mHealth, a non-profit organization, has been building and promoting common data schemas and application programming interfaces (APIs) to expand the scale, power, interoperability, and flexibility of mobile health.

With the global wearables market projected to triple in value between 2016 and 20211 and with the adoption of “mobile health” in NIH projects accelerating, now is the time for the mHealth ecosystem to adopt common data and API standards.

Methods

The Open mHealth platform uses JSON REST APIs. Open mHealth schemas are written in JSON Schema draft 4. Our initial schema modeling was driven by use cases in the domain of preventive cardiology, which needs data elements collected by a wide range of consumer-facing apps and devices (e.g., physical activity, step count, blood pressure, body weight) that are needed for many other clinical conditions. We also identified modeling requirements concerning the representation of time, measurement units, provenance, and data access. To address the use case of medication adherence, we organized a Clinical Measures workgroup (CLIME) with participation from two device manufacturers, a medication adherence researcher, and a pharmacist to model the medication cycle from prescription to dispensing to individual dose taking by the patient.

Additional drivers of schema development included our participation in the NIH-funded Center of Excellence for Mobile Data to Knowledge (MD2K)2, which models mobility patterns (from geolocation, accelerometry, etc.) and psychophysiological measures (e.g., stress from heart inter-beat interval, inspiratory and expiratory time), among several others. We modeled sleep episode and sleep duration, as sleep is an important measure for managing a number of chronic conditions. We are testing and supporting use of Open mHealth’s open source tools by a number of organizations, as well as by mobile research studies based on ResearchKit and ResearchStack.

Results

To date, we have defined more than 30 “generic” schemas on physical measurements and their units, time, descriptive statistics (e.g., average, median) and other value sets (e.g., temporal relationship to meal, to sleep, to physical activity); and some 60 “clinical” schemas for data like body weight and blood glucose that reference generic schemas as needed (e.g., for physical units). Schemas reference SNOMED or LOINC codes at the NCBO BioPortal terminology server. Schemas are intentionally parsimonious on what data is required (e.g., the blood glucose schema models and allows, but does not require, metadata on the temporal relationship to meals).
To facilitate data access and integration, we have developed Shimmer\(^5\) an open source application for accessing mHealth data in Open mHealth format from common sensor makers. Shimmer comprises individual shims, a resource server, and a console. A shim is a library that can communicate with a specific third-party API (e.g., Fitbit, Moves, Withings), handling authentication, requests for data and mapping of that data into Open mHealth format. A shim generates data points plus header information such as date of creation, acquisition provenance and data source.

The resource server exposes an API to retrieve data points. The server handles API requests by delegating them to the correct shim and manages third-party access tokens on behalf of shims. The console provides a simple web interface that helps users interact with the resource server. We have also developed Granola, which allows access to HealthKit data on iOS and exports them in Open mHealth-compliant schemas to the datastore of choice. Finally, we have defined a set of visualization modules for common clinical data types (like blood pressure, heart rate, weight).

Open mHealth schemas, Shimmer, Granola and visualization modules are publicly available on the Open mHealth GitHub repository\(^6\) with detailed documentation. Additional documentation is available at openmhealth.org. Community members can adapt and evolve the code to customize it for their specific needs. Open mHealth tools had over 14,000 downloads (11,000 in the last 15 months), 155 forks and 425 GitHub stars, and are being used by organizations including Merck, Kaiser Permanente, Stanford Hospital, University of Toronto, the Nurses Health Study, UC Davis, and Columbia Medical Center.

**Discussion**

Open mHealth’s schemas and open API specification are designed to allow and encourage, but not require, the ideal formatting and description of mHealth data for supporting clinical and self care. This flexibility allows different clinical domains to evolve towards pragmatic standards at their own pace, while still promoting data interoperability.

Current work includes: community engagement towards approval as an IEEE standard through Open mHealth’s IEEE P1752 Working Group\(^7\); more detailed modeling of metadata including provenance (as part of the NSF-funded mProv project); participation in complementary efforts, like the Consumer Technology Association Working Group on Consumer Stress Technologies and the IEEE Wearables & Medical Interoperability Workshop; collaborative development of additional schemas on sleep, symptoms (such as pain), and patient-reported outcomes (PROs) with clinical, device and academic partners; and development of an Open mHealth-based mFHIR Implementation Guide and associated FHIR extensions through partnerships with the HL7 FHIR community and a new mHealth track in the HL7 Partners in Interoperability program. Open mHealth’s goal is to enable the integration and use of personally generated data alongside clinically generated data from EHRs and other health information technologies.

**Acknowledgements**

Open mHealth is a project of the Tides Center, and is funded in part by NIBIB grants 1R24ES028492-01 and 1U54EB020404 (through funds provided by the trans-NIH Big Data-to-Knowledge (BD2K) initiative) and by ACI-1640813 (from NSF).

**References**

Exploring Hadoop-Based Data Lakes for Research
Rajan Chandras¹ MS MSc PMP, Michael N. Cantor¹,² MD MA, Jeff Shein¹ BA
NYU Langone Health¹, Departments of Medicine and Population Health, New York
University School of Medicine², New York, NY USA

Introduction
An important aspect of research data governance is the provisioning and management of data for research, analytics and data mining. In an academic medical center, such analytic requirements typically span multiple “dimensions” of data such as type, source and format, and often do not fall within conventional analytic patterns for a variety of reasons to be discussed below. At NYU Langone Health, we explored multiple architectural alternatives as well as architectures using “big data” approaches and the Hadoop® technology ecosystem in order to address selected research and enterprise analytic requests. This paper discusses the early successes and challenges we have faced and the lessons learned in our journey towards implementing an enterprise data lake (1, 2, and as described below).

Methods
1. Data Management Platform: Data required for research in an academic medical center can be very heterogeneous and voluminous (2), varying by type of data (clinical, financial, social, genomic, etc.); source of data (internal clinical and non-clinical application systems, “internet of things” (IoT) data from wearables and devices, external data such as from CMS or the state, etc.); and data format (e.g. relational, text files, spreadsheets, images, audio, video). We chose the Hadoop® ecosystem for its ability to store and process all such types of data. (3)

2. Architectural Alternatives: Conventional approaches are inadequate to address “big data” challenges (4). We considered and eliminated several other architectural approaches to provisioning researchers with their data requirements (the “end users”), for the reasons stated in Table 1 below:

<table>
<thead>
<tr>
<th>Approach</th>
<th>Reasons for Elimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open access to data sources</td>
<td>Providing access to all users to all data sources is not practical and not secure</td>
</tr>
<tr>
<td>Conventional data warehouse</td>
<td>Requires complex data model possibly from vendor; expensive to build and maintain</td>
</tr>
<tr>
<td>Data virtualization</td>
<td>a.k.a. Enterprise information integration (6): Does not scale; cost escalates quickly</td>
</tr>
<tr>
<td>Conventional databases</td>
<td>E.g. Oracle: Requires data modeling and transformation; cannot handle different types of data (e.g. images) elegantly; licensing can be relatively expensive for very high data volumes; does not scale to petabyte-level requirements</td>
</tr>
</tbody>
</table>

3. Implementation: The Hadoop® ecosystem offers a plethora of technologies for data ingestion, storage, processing and provisioning (6, 7). We attempted to select an appropriate set of technologies for each use case, keeping in mind overall architectural cohesion and supportability. For example, using Sqoop for data ingestion, Parquet for data storage, Spark for processing and Hive/Impala for querying ensures that the tools used are foundational, generic and integrate with Sentry for data security, while the data itself can be accessed using a broad set of languages such as SQL, R and SAS. A “lift and shift” approach was used to ingest data with little or no data transformation. Users are required to learn source data structures, relationships and constraints.

Results
We successfully implemented the use cases listed in Table 2 below using Hadoop, the first two as proof-of-concepts that may be re-implemented as full solutions, and the last three as production solutions.

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Data Source</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Analytics</td>
<td>New York State SPARCS data</td>
<td>Pig, HDFS, Hive, Impala, Tableau</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>Facebook®, Twitter® and Instagram® data</td>
<td>MongoDB, Flume, HBase, Pig, DropWizard Framework, AMCharts UI</td>
</tr>
<tr>
<td>Clinical Notes NLP for Studies</td>
<td>EHR data</td>
<td>Sqoop, HDFS, R</td>
</tr>
</tbody>
</table>
In addition, we are currently in the early stages of ingesting a vast amount of data spread across clinical data 
“domains” such as demographics, diagnoses, procedures, encounters, surgery etc. as well as non-clinical data (e.g. 
claims) from internal and external data sources into Hadoop, to be enriched with metadata and reference data, 
governed through technology and standard operating procedures, and provisioned to authorized users. We expect the 
collective result of these efforts to result in an enterprise data lake: an untransformed but well organized, governed, 
content- and metadata-rich self-service data repository that serves the purpose of clinicians, analysts and researchers.

Discussion

Hadoop-based data lakes that implement “lift-and-shift” (i.e. non-transforming), late binding (to analytic models), 
self-service data architectures are a natural fit for research analytics for a number of reasons: Researchers usually 
require broader and deeper data, not limited to specific views or subsets of data; they prefer the data to be in source 
format so that it is easier to join and process differently for different analytic studies; they possess far more hands-on 
data analysis and programming skills (SQL, SAS, R, Python) than typical business users; they prefer a “self-service” 
model rather than depending on IT for their data requirements; and their requirements are changing (e.g. study- 
based) and evolutionary (as they expand their analysis). The Hadoop platform supports these requirements well, 
because it does not enforce structure (as relational databases do), allows unrelated data sets from multiple sources to 
coexist in a single integrated environment, supports different types of data, and can be scaled as needed. 

We have found that selecting the right Hadoop components for data ingestion and storage are very 
important, such as using the correct data storage format (e.g. Parquet). Use of NoSQL databases should be 
considered with care as they can add avoidable complexity. Integration with Sentry for data security and metadata 
gathering is important. Detailed analysis is needed to assess each potential Hadoop use case, e.g. the types and 
sores of data, need for data refresh, whether it has PHI, and how it will be accessed. Hadoop is a relatively new 
technology without established methodologies or abundant, inexpensive skills, and there will be a learning curve for 
the implementation team as well as users. Last but not least, selecting the right Hadoop distribution can be 
important, although few organizations have the expertise, time or budget to conduct a detailed comparative 
assessment of different distributions. 

Above everything else, there must be architectural vision and oversight that leads towards a single, unified, 
cohesive data fabric, and ensures that individual initiatives do not devolve into separate “data puddles” or deteriorate 
into a “data swamp”. At NYU Langone Health, we are striving to accomplish this by ensuring that requirements are 
fit for Hadoop and the data lake; a relatively small set of established Hadoop components is used; data ingestion and 
provisioning is centrally governed; and data access is granted selectively. We have also embarked on a parallel 
initiative to govern metadata and reference data that will integrate with the data lake, thus enhancing the usability 
and business value of the data lake.

References

2. Krzysztof J. Cios, G. William Moore. Uniqueness of medical data mining. Artificial Intelligence in Medicine, 
Volume 26, Issues 1–2, September–October 2002 [Link]
approach. PMID: 25991244 [Link]
4. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. PMID: 25825667 [Link]
12-401696-5. [Link]
Predicting Postoperative Bleeding Risk for Patients Undergoing Colorectal Surgery Using Machine Learning

David Chen, PhD¹, Naveed Afzal, PhD¹, Sunghwan Sohn, PhD¹, Elizabeth B. Habermann, PhD², James M. Naessens, ScD², David W. Larson, MD, MBA³, Hongfang Liu, PhD¹

¹Division of Biomedical Statistics and Informatics, ²Division of Health Care Policy and Research, ³Division of Colorectal Surgery, Mayo Clinic, Rochester, MN

Introduction
Postoperative bleeding (POB) is a fairly rare but serious complication following colorectal surgery (CRS), occurring in between 2-14% in all CRS cases¹². Furthermore, POB is often associated with other complications including ileus, anastomotic leakage, and reactions to transfusion of allogeneic blood components. There is a lack of consensus for risk factors for POB, suggesting the complexity of the underlying cause and lack of standardization of practice. Machine learning (ML) has been touted as a way to learn models that represent difficult non-linear systems, particularly when utilizing high-dimensional “big data”. The objective of this study is to present a predictive ML model for identifying patients at high risk of POB using data extracted from electronic health records (EHR).

Methods
A retrospective review of adult patients who underwent CRS as the primary procedure was conducted at Mayo Clinic Methodist Hospital between January 1998 and December 2014. POB was determined through ICD-9 code for any type of hemorrhage. A total of 12,402 unique patients undergoing 13,399 CRS procedures were found. Of the 13,399 cases, 1,680 cases (12.5%) were found to have experienced POB. One hundred and seventeen variables were extracted from EHR (Table 1). Patient provided information, symptoms, physiological values, observational factors, and terms extracted from clinical notes were limited to a window of 7 days prior to surgery. Demographics and comorbidities were limited to data up to 3 months prior to surgery. Temporally varying physiologic measurements were abstracted into 6 scaler elements (minimum, mean, and maximum values and the minimum, mean, and maximum change in values) to extract possible information contained in temporal dynamics of the signal. The first measurement post-operation of temporal features (e.g. first postoperative systolic blood pressure measurement) was also included. A total of 78 scalar variables were created from the 11 time-varying physiological measurements. In addition to the structured variables, another 16 variables were generated using MedTagger, an open-source tool which enables discretization of data that are locked in unstructured clinical notes (http://ohnlp.org/index.php/MedTagger). A total of 299 variables were created for the final dataset. All variables that were missing more than 50% of data were excluded. The remaining missing data was imputed using the “MissForest” package in R. Compared to complete case analysis, using imputation has been shown to yield more accurate models and better fit due to more accessibility to a greater number of observations and better representation of the true distribution of the variables³.

Logistic regression (LR), and gradient boosting machine (GBM) methods were used to predict POB. GBM was used due to its robustness to missing values and good out-of-box performance. Evaluation of each model was performed using 10-fold cross-validation repeated 10 times. Due to the imbalanced class distribution, the observations with positive POB of the training set were oversampled while negative observations were undersampled such that each training set had a near 50/50 balance. Area under the curve (AUC) is used to assess the performance of the models. The Boruta method was used for variable selection. The method iteratively compares variable importance of given variables to shadow variables on a random forest model. Variables with significantly higher importance are kept while non-important variables are discarded. Following Boruta variable selection, the model was further simplified by empirically reducing the number of predictors until AUC is significantly lowered. All models were trained and evaluated in R.

Results and Discussion
Table 1 shows the variables extracted from EHR. Boruta variable selection found age, bmi, procedure, surgical length, laparoscopic, intraoperative complication, anemia, embolism, sepsis, ulcer, inflammation, injury, weakness, preoperative bleeding, malignant colon cancer, malignant rectal cancer, malignant other cancers, heart failure, coronary artery disease, kidney disease, colitis, hemophilia, activity level, fall risk score, mobility, nutrition, pain,
diastolic blood pressure, systolic blood pressure, heart rate, O₂ saturation, respiratory rate, temperature, hemoglobin level, leukocyte level, glucose, platelet, creatinine, potassium, sodium, alkaline phosphatase, and NLP abstracted abscess to be significant predictors of POB. Table 2 shows performance of various different models. GBM outperforms LR significant.

Table 1: Variables abstracted from EHR. †Variables used for LR model. ¥Variables significant for ML model.

<table>
<thead>
<tr>
<th>Variable Types</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>age, gender, BMI, race</td>
</tr>
<tr>
<td>Patient Provided Information</td>
<td>anesthesia complication, bleeding complication, difficulty with pain, difficulty dressing, difficulty eating, difficulty walking, difficulty taking medications, alcohol usage, smoking, drug usage, anticoagulant medication</td>
</tr>
<tr>
<td>Symptoms</td>
<td>weak, surgical site infection, bleed, transfusion, wound disruption, shock, non-healing wound, abdominal pain, genital inflammation, sepsis, ulcer, inflammation, angina, hernia, injury, heart rhythm, nausea, vertigo, anemia</td>
</tr>
<tr>
<td>Comorbidities</td>
<td>cancer, heart failure, hypertension, coronary artery disease, malnutrition, cardiomyopathy, embolism, kidney disease, diabetes, diverticulitis, hepatitis, colitis, cholecystitis, cholangitis, pancreatitis, nephritis, cystitis, hemophilia</td>
</tr>
<tr>
<td>Physiological Values</td>
<td>diastolic blood pressure, systolic blood pressure, heart rate, O₂ saturation, respiratory rate, temperature, hemoglobin, leukocyte, neutrophil, basophil, lymphocyte, monocyte, glucose, platelet, hematocrit, creatinine, potassium, sodium, aspartate aminotransferase</td>
</tr>
<tr>
<td>Observational Factors</td>
<td>ASA score, respiratory pattern, nutrition, orientation, fall risk score, mobility, pain</td>
</tr>
<tr>
<td>Operational Variables</td>
<td>surgery length, procedure, surgical approach, intra-operational complication</td>
</tr>
</tbody>
</table>

Table 2: Performance of various models. P-value (in parenthesis) is with reference to all variables.

<table>
<thead>
<tr>
<th>Subset</th>
<th>All</th>
<th>Boruta</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.720 (1.00)</td>
<td>0.735 (0.35)</td>
<td>0.672 (&lt;0.01)</td>
</tr>
<tr>
<td>GBM</td>
<td>0.819 (1.00)</td>
<td>0.822 (0.72)</td>
<td>0.752 (&lt;0.01)</td>
</tr>
</tbody>
</table>

Surveillance for POB is a resource intensive task due to need for human attention and laboratory testing. Delay in identifying a bleeding event can lead to significantly worse outcomes, longer hospital stays, and increased costs. Early and specific surveillance can improve resource allocation, particularly if early intervention can reduce the amount of transfused blood. However, bleeding events are difficult to predict in part due to the inherent variability of any underlying causes. ML methods can often do a better job of modeling these systems because the models do not assume linearity. In particular, ensemble models such as GBM can do a good job of approximating a highly variable system by systematically predicting harder to predict observations. In this way, ensemble methods can be examined to identify variables which may not be readily important in traditional feature selection. In this work, the GBM model found nutrition, heart failure, kidney disease, and fatigue to be important predictors of POB.

Conclusion
The results presented in this work demonstrate the utility of ML methods in predicting POB following CRS procedures. New risk factors such as mobility issues, weakness, sodium, and fall risk score were identified which could lead to improved understanding of causes of POB and to better allocate resources for POB surveillance.

References
Executable Data Science Notebook for Phenotype Discovery Pipeline Using Observational Data
Robert Chen, BS1, Jon Duke, MD1, Jimeng Sun, PhD1
1School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA

INTRODUCTION: Alongside the emergence of clinical data research networks (CDRNs), such as Observational Health Data Sciences and Informatics (OHDSI) consortium,[1] computational phenotyping from electronic health records (EHR) data has become a major interest due to its potential for use cases in precision medicine, population health management, and in the understanding of complex diseases such as heart failure.[2] Researchers typically approach phenotyping with diverse workflows which may lead to irreproducible results and difficulty with widespread adoption. However, for advancements in medical care, it is paramount for data scientists to perform tasks in a structured manner that is consistent between people in order to preserve scientific validity. To address this shortcoming, we designed and implemented, a modular and executable phenotyping pipeline in the form of an easily portable data science notebook and showcased it on use cases for phenotype discovery and phenotype assignment for heart failure patients. Our study showcases a proof of concept of an open-source, scalable and modular phenotyping framework, using the lab notebook form factor with which many medical data scientists are familiar. It is the first such implementation of a data science analytics environment for performing phenotyping tasks on datasets that may be obtained in the OMOP common data model and should motivate future initiatives on releasing share-able, actionable code samples that can be easily customized to any medical data scientist’s preferences.

METHODS: The phenotyping pipeline is an interactive coding environment that allows the user to view and execute code to perform all relevant tasks required for extracting phenotypes from source data. The pipeline is implemented in a Jupyter Notebook. An electronic health record database that follows the Observational Medical Outcomes Partnership common data model (OMOP CDM) is read into the phenotyping pipeline via a database connector. The phenotyping pipeline is implemented in a lab notebook and executes a series of tasks that are required for extraction of phenotypes from the source data: data transformation, phenotype extraction, follow-up analysis, displaying results in-line, and exporting final phenotype definitions and cohorts as well as the discovery and assignment pipeline.

Use Cases Demonstrated: We demonstrated 3 case studies with our pipeline: 1. Discovering new phenotypes on a new cohort with a pipeline, 2. Assigning patients from a new cohort into phenotypes computed with a previous cohort, 3. Customization of the pipeline with a new phenotyping algorithm.

RESULTS: To demonstrate the usage of the pipeline in real-world phenotyping tasks, we performed 3 case studies. The first use case is the implementation of the pipeline on a cohort of heart failure patients in the Sutter Palo Alto Medical Foundation clinic, and the subsequent usage of the pipeline with the same configurations for discovery of heart failure subtypes on a cohort of patients from the CMS DE-SynPUF dataset. The second use case is the assignment of patients from the CMS DE-SynPUF cohort to the phenotypes obtained from the Sutter PAMF cohort. The third use case is the customization of the notebook with a different phenotyping algorithm. In all 3 case studies, the phenotype definitions are saved in JSON format and the pipeline and associated system configurations are stored.

Figure 1. Overview of the workflow and phenotyping pipeline components.

Phenotyping Lab Notebook Details: First, a database connector is required in order to retrieve input data from an OMOP database (figure 2A). Next, in the data transformation step, features from various OMOP concept tables are pulled from the database and transformed into the required data structures for phenotype extraction. In the phenotype extraction step, the phenotyping algorithm is run and results are stored in a JSON object in the results directory. In our specific example, non-negative tensor factorization (Marble) algorithm is used for phenotype extraction, and resulting phenotypes are displayed in the notebook (figure 2B).[3] Afterwards, follow-up analysis can be performed. In our example, survival analysis via Kaplan-Meier analysis is conducted (figure 2C). Finally, the results of the phenotype discovery pipeline are saved and exported into a common results directory. Two JSON objects are stored: one for the phenotype definitions discovered, and one for the pipeline itself, which can be launched by others in order to repeat the specific phenotype discovery task.
Figure 2: Screenshots from the phenotyping pipeline notebook. A) specifications for a connection from the lab notebook to a database that follows the OMOP Common Data Model. B) excerpts from the code for running the non-negative tensor factorization based phenotyping algorithm. Descriptions of the phenotypes are displayed in-line. Display of general statistics of the phenotype results are displayed in bar-chart format inline (with Matplotlib). C) excerpts of code for follow-up analysis that can be conducted on the phenotypes. Kaplan-Meier analysis is shown with Kaplan-Meier curve and differential survival shown in bar chart format.

together in a shared directory. In all three case studies, a phenotype is defined as a set of patients and associated ICD-9 codes most relevant to that phenotype.

**Case Study 1:** We used the pipeline to discover phenotypes for 4,370 heart failure patients in the Sutter Health dataset. Features pertaining to conditions (e.g., ICD-9 codes) from a 2-year observation window were extracted and transformed to tensors, which are fed into the phenotyping algorithm. Phenotypes extracted are stored in JSON objects.

**Case Study 2:** We use the phenotyping assignment pipeline to assign heart failure patients from the Sutter Health dataset to the phenotype definitions obtained from the CMS DE-SynPUF dataset. To perform this, patients from the Sutter Health dataset are projected onto the space of phenotypes extracted from the CMS SynPUF dataset, using the assignment weights stored in the phenotype definitions from the discovery process via the lab notebook.

**Case Study 3:** We demonstrate the modularity of the lab notebook by modifying it to use a K-means algorithm for determining phenotypes, in place of the original non-negative tensor factorization algorithm, and using K-means, followed by recursive feature elimination as an algorithm to discover 10 phenotypes of heart failure patients from the Sutter Health dataset.

**DISCUSSION:** Standardization and portability of practices allows for ease of collaboration, validation and reproducibility of results in the healthcare analytics practices involving electronic health record data. Although we showcased the usage of our phenotyping pipeline notebooks for subtyping of heart failure patients, our study was not intended to be a robust discovery and validation process for heart failure phenotypes. Future research should leverage standardized phenotyping analytic pipelines such as ours, in an effort to iteratively identify and validate phenotypes across various patient cohorts of interest. While our pipeline currently integrates a tensor factorization based phenotyping algorithm, it is important to note that our framework proposes a standardized workflow for any unsupervised phenotyping algorithm. Finally, future work should focus on implementations of the pipeline in other settings such as Zeppelin notebooks. Our work should motivate future studies that leverage standardized phenotyping analytic pipelines such as ours, in an effort to iteratively identify and validate phenotype definitions for subtypes of highly complex diseases such as rheumatoid arthritis, diabetes, and cardiovascular disease.

**CONCLUSION:** We proposed and implemented a modular and executable data science pipeline for performing phenotype discovery tasks via an easily sharable notebook environment. We applied this pipeline successfully to several case studies, all of which illustrated the portability and customizability of the pipeline.

**References**


Interoperability between Value Sets for Clinical Research and Healthcare: Mapping Value Sets between the Clinical Data Interchange Standards Consortium (CDISC) and Meaningful Use

Raja A. Cholan, BS¹, Olivier Bodenreider, MD, PhD²
¹Oregon Health & Science University, Portland, OR; ²National Library of Medicine, Bethesda, MD

Introduction: Historically, there have been substantial – but, independent – efforts to standardize medical information for clinical research data and those for healthcare data. For instance, in the biomedical research domain, the Clinical Data Interchange Standards Consortium (CDISC) was developed for data exchange in clinical trials used by the Food and Drug Administration and pharmaceutical companies. CDISC uses codes and terms from clinical terminologies from the National Cancer Institute (NCI) thesaurus. The Value Set Authority Center (VSAC) maintains value sets for clinical quality measures used in value-based programs like Meaningful Use (MU). The VSAC value sets use codes and terms from standard terminologies (e.g., SNOMED CT, RxNorm, LOINC) found in most Electronic Health Record (EHR) systems.

With EHR adoption becoming nearly ubiquitous, researchers want to leverage EHRs to conduct pragmatic trials in which they can counteract recruitment challenges, burdensome data collection, and uncertain generalizability of results. Therefore, there is a need for interoperability between healthcare data (VSAC) and clinical research data (CDISC). In previous work, we and others have studied value sets found in the VSAC. Within biomedical research, meanwhile, there is a lack of methodological literature about value set engineering, particularly in CDISC controlled terminologies. In this work, our objective is to assess semantic interoperability between CDISC value sets and MU value sets. Our study questions were (1) What are the semantic characteristics of the concepts used in CDISC and VSAC value sets?; (2) To what extent do existing value sets in the VSAC represent value sets in CDISC?; (3) Can we create a surrogate source of value sets that wouldn’t already exist in the VSAC – by using the Unified Medical Language System (UMLS) Metathesaurus to represent CDISC value sets from standard terminologies?

Methods: We sought to identify gaps and similarities between clinical research value sets and healthcare quality value sets. First, we gathered the lists of value sets from CDISC (using CDISC’s FTP website) and those from VSAC (using the VSAC API). Then, we mapped the codes from each source to a concept unique identifier (CUI) from the UMLS (using the UMLS API). Then, for the CUIs representing the CDISC value sets, we created surrogate value sets using associated atoms from standard terminologies in the UMLS (using the UMLS API). Next, we used the Jaccard index (intersection/union) to compare the overlap between the terminologies. Finally, we evaluated the alignments obtained by the two sources to determine what value sets and concepts in CDISC were not covered by value sets and concepts in standard terminologies.

Results:

1. Value Set Semantic Profiles. The distribution of semantic groups found in CDISC concepts is significantly different from that in VSAC concepts. The CDISC value sets mostly represent administrative concepts, and a small subset of procedures and disorders. The VSAC value sets mostly represent clinical concepts, such as disorders, drugs, procedures, and very few administrative concepts. Figure 1 displays the value set semantic profiles in CDISC and VSAC.

![Figure 1. CDISC and VSAC Semantic Profiles](image)
2. Coverage of CDISC by VSAC. ~92% of CDISC codes could be mapped to a UMLS CUI. Sampling the 8% not mapped shows provisional codes added to NCI thesaurus in 12/2016. In VSAC, meanwhile, 99.8% of the source codes were mapped to a UMLS CUI. We computed the coverage of CDISC by VSAC by analyzing the count of CDISC value sets that shared at least 1 UMLS CUI with a corresponding VSAC value set. 114/643 (17.7%) CDISC value sets share at least 1 UMLS CUI with a VSAC value set with a mean Jaccard index of 0.028.

Figure 2. Coverage of CDISC by VSAC

3. Surrogate Value Set Coverage. We were able to create surrogate value sets to cover CDISC value sets from SNOMED CT, LOINC, ICD-10-CM, ICD-9-CM, CPT, RxNorm, HCPCS, and ICD10PCS terminologies. The surrogate value sets comprised of SNOMED CT codes covered 178 (28%) CDISC value sets with a mean Jaccard index of 0.37, and LOINC surrogate value sets covered 130 (26%) CDISC value sets with a mean Jaccard index of 0.29. See Table 1 for a full list of the number of CDISC value sets covered by each standard terminology, and the number of UMLS CUIs from CDISC value sets covered by each surrogate value set. The counts of surrogate value sets represent value sets which covered an existing CDISC value set for at least 1 UMLS CUI.

Table 1. Counts of CDISC value sets covered by surrogate value sets from standard terminologies

<table>
<thead>
<tr>
<th>Terminology</th>
<th># of CDISC Value Sets Covered (%)</th>
<th># of UMLS CUIs covered in CDISC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNOMED CT</td>
<td>178 (27.7%)</td>
<td>4315 (26.4%)</td>
</tr>
<tr>
<td>LOINC</td>
<td>130 (25.7%)</td>
<td>1271 (7.8%)</td>
</tr>
<tr>
<td>ICD-10-CM</td>
<td>15 (2.3%)</td>
<td>119 (0.7%)</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>14 (2.6%)</td>
<td>84 (0.5%)</td>
</tr>
<tr>
<td>CPT</td>
<td>10 (1.6%)</td>
<td>227 (1.4%)</td>
</tr>
<tr>
<td>RxNorm</td>
<td>7 (1.1%)</td>
<td>126 (0.8%)</td>
</tr>
<tr>
<td>HCPCS</td>
<td>2 (0.3%)</td>
<td>5 (0.03%)</td>
</tr>
<tr>
<td>ICD10PCS</td>
<td>1 (0.2%)</td>
<td>1 (0.01%)</td>
</tr>
</tbody>
</table>

Conclusions: VSAC/MU value sets mainly cover clinical concepts of interest such as diagnoses, drugs, procedures, and not many administrative concepts. CDISC value sets, meanwhile, essentially cover administrative concepts, and a small subset of disorders and procedures. Only about 18% of CDISC value sets can be represented by a VSAC value set with a low mean Jaccard index. Surrogate value sets created from SNOMED CT could only represent 28% of CDISC value sets, while LOINC surrogate value sets represented 26% of value sets; both SNOMED CT and LOINC surrogate value sets had higher (but less than ideal) Jaccard scores than those found in the coverage by VSAC value sets.

The overall interoperability between CDISC and MU was limited at best. Interestingly, there are a number of value sets for questionnaires, functional assessments, experience scales etc. in CDISC with little or no coverage by LOINC or SNOMED CT. One suggestion is for LOINC and SNOMED CT to look into these and include them in future versions, if appropriate.
Personalized Vital Sign Modeling to Predict Health Status

Jessilyn Dunn, PhD1, Lukasz Kidzinski, PhD1, Ryan Runge2, Scott Delp, PhD1, Michael Snyder, PhD1, Trevor Hastie, PhD1
1Stanford University, Palo Alto, CA, USA; 2UC Berkeley, CA, USA

Introduction

A routine clinical examination consists of vital signs measurement and blood and urine testing (referred to here as clinical laboratory tests) to evaluate overall health status and risk for illness, to detect any abnormalities such as infection or existence of a chronic illness, and to develop individual healthcare management strategies. These tests require an individual to come into the clinic, often for more than one visit for laboratory sampling and for clinical interpretation by the physician. While vital signs are generally non-specific to a diagnosis, they are highly useful for assessing overall health and triaging patients rapidly in both the routine clinic and emergency settings. Automated methods to directly relate vital signs to health status would assist in diagnostic testing and medical decision making both within and outside of the clinic. In this study, we sought to precisely understand the relationship between vital signs and health status, as well as the individual variability in this relationship towards the practice of personalized medicine. We focus on vital signs because, of those measurements taken in a traditional clinic visit, the vital signs are currently the only measurements that can be continuously, passively, and remotely monitored outside of the clinic.

Methods

We explored clinical records from 28,694 patients at Stanford Hospital, containing 31,543,209 laboratory test results (18,360,532 from our 50 labs of interest) and 885,966 vital signs measurements (552,145 pulse measurements and 333,821 temperature measurements). The patient population contained ~10,000 each of diabetic, prediabetic, and normoglycemic individuals. There were 2,662,029 clinical laboratory tests that occurred on the same day as both a pulse and temperature measurement.

We chose 50 clinical laboratory tests (e.g. Blood Glucose, Platelet Count, etc.) based on their diagnostic utility in a primary care setting, and calculated the coefficient of multiple determination (R^2 value) for each clinical laboratory test using a bivariate linear regression model (clinical laboratory test ~ pulse + temperature). We ranked the clinical laboratory tests by their R^2 value and F-statistic and chose the top 4 clinical laboratory tests for further analysis. We binned each clinical laboratory test into 40 bins, calculated the mean pulse or temperature for each bin, and examined linear, quadratic, and spline model fits.

Results

For the majority of the clinical laboratory tests, the relationship between the bin mean and vital signs were best described by a quadratic model (Figure 1). Bins for neutrophils, lymphocytes, albumin, and procalcitonin contain 1775, 1818, 1941, and 2243 clinical measurements per bin, respectively. We found that at the individual patient level, there was significant variability in the relationship between the clinical laboratory values and the vital signs as demonstrated by the wide distribution of linear mixed model slopes (Figure 2). This finding underscores the importance of personalized baselines as key factors in models involved in assisted medical decision making.

Discussion

Here, we show that vital signs that are measured in the clinic can be used as a basis to predict health status (e.g. clinical laboratory tests), and that models using individualized baselines can improve such predictions. In current practice, clinical measurements are collected intermittently. Continuous, longitudinal, and remote monitoring using wearable devices can be used in the future to improve the accuracy of the models developed herein. Further, our finding that personalized models often diverge from population models emphasizes the need for personalized
medicine, or ensuring that each medical decision is tailored to the specific individual rather than the population mean. Overall, we establish population-level models relating clinical labs and vital signs, and highlight that information provided by repeated measurements of vital signs within an individual can be used to improve these models toward personalized care. In the future, the models we develop here may be used to predict health status using only remotely collected vital signs data from wearable devices.

Figure 1. Quadratic models of the relationship between vital signs and clinical laboratory tests. Each point represents the mean pulse (A) or body temperature (B) for the corresponding clinical laboratory test bin.

Figure 2. The distribution of slopes from linear mixed models of individual patients for the relationship between either pulse (A) or temperature (B) and the clinical laboratory tests.

References
A Transparent System for Evaluating Data Completeness in the Accessible Research Commons for Health (ARCH) Network

Hossein Estiri, PhD1,2,3, Jeffrey G Klann, PhD1,2,3, Sarah R Weiler, PhD1, Kenneth D Mandl, MD, MPH1,4, Shawn N Murphy, MD, PhD1,2,3

1Harvard Medical School, Boston, MA; 2Massachusetts General Hospital, Boston, MA; 3Partners Healthcare, Boston, MA; 4Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA

Introduction

The quality of data in federated research data networks is paramount, and yet difficult to maintain and monitor, given a distributed and potentially diverse set of sites, stakeholders, information systems, and resources.1 Investigators using the network rarely can access a transparent data quality reporting system.2 Here, we describe the approach we have taken in the Accessible Research Commons for Health (ARCH) Clinical Data Research Network (CDRN) – formerly known as Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS)3 – for developing technology and establishing a workflow to evaluate data completeness in EHR repositories across its partner sites. This data quality system enables the ARCH network to monitor, and if needed, adjust its data transformation processes in order to maintain high quality data for secondary use. We also discuss implementation of the system in seven ARCH clinical sites between February and August 2017.

Methods

The ARCH data quality workflow is comprised of two processes, in which outputs of the first process are routed back as inputs of the second process as a feedback loop (Figure 1). This feedback system is initiated using two open source tools. Through the first process, participating ARCH sites generate standard reports on the completeness of their respective clinical data repositories, using DQ-e-c, an open source tool that evaluates completeness and conformance in EHR data repositories. ARCH partner sites maintain the latest install of the tool in their local environments and can participate in refining DQ-e-c through GitHub. Every run of the tool at a given site produces a number of comma-separated values (CSV) files that embody aggregated results. ARCH partner sites are encouraged to run the tool after all major data reloads that happen quarterly across the network. Sites share two of the output flat files with the ARCH data quality team for a network-level analysis and providing feedback.

* The first process is illustrated on the left of the diagram. The second process (on the right) illustrates the feedback loop. A snapshot of ARCH Vue interactive dashboard pages are also provided as a product of the second process.

**Figure 1.** ARCH workflow and technology to evaluating completeness in the network’s distributed clinical data.
The second process is production of network-level aggregate report and a set of site-level feedback reports. Data processing and analyses for producing these reports are conducted by the ARCH data quality team, using a second open source tool, Vue, which uses DQc report files to compile: (1) an interactive network-level dashboard, ARCH-Vue dashboard, intended for the ARCH governance committee, and 2) a series of site-level reports that are shared with each individual partner site. The site-level reports enable ARCH partner sites to assess results of their completeness indices in the context of other partner sites and the entire network. After accumulating the DQc-report files, Vue re-calculates completeness/missingness percentages by aggregating the raw numbers, and uses test dates available in DQc-report files to visualize the latest completeness status, as well as a longitudinal overview of completeness at both site and network levels to show variations in completeness metrics over time.

Results

Vue outputs are categorized into four sections. First, a preview of unique counts recaps the number of unique patients, encounters, diagnoses, and procedures. The preview is most useful in the network-level dashboard, where running DQc is virtually impossible, due to the federated nature of network. Changes in unique counts over data loads are also visualized in the preview. Second, Vue visualizes completeness across network by table, which in our current implementation at ARCH network are tables in PCORnet Common Data Model. Column-level missingness are presented in the third section for each table – DQc quantifies missingness by column, where missingness is measured in the availability of NA/NULL cells or “nonsensical” characters, such as an exclamation point (!) or a hashtag (#). Finally, Vue re-compiles a network-level aggregation of missingness in key indicators; a growing list of indicators such as percent of patients missing records on race/ethnicity, blood pressure, medication, weight, and height. Vue visualizes the latest status of missingness in key indicators and also provides site-level changes over time. These visualizations are also embedded in the interactive ARCH Vue dashboard (Figure 1).

Discussion

The Accessible Research Commons for Health’s technology and workflow for evaluating data completeness across its partner sites enables the network to monitor, and if needed, adjust its data management processes in order to maintain high quality data for secondary use. To implement this system at any distributed clinical data network, the only action that needs to be taken by partner sites in is to run DQc and share their results with the network’s governing board. Between February and August 2017 seven of the ARCH partner sites have installed and run DQc at least once. Running Vue on aggregate site-level outputs is almost equivalent of running DQc on a dataset comprised of merging all distributed data repositories across the network, which makes Vue extremely efficient from data governance and computation for multi-site distributed clinical data networks. Network-wide implementation may lead to increased participation in this system. Although this system requires minimal data-sharing, we are evaluating potential data security concerns that may hinder higher partner site participation in the system. Finally, the Vue dashboard may be repurposed as a data-profiling tool, enabling data consumers (e.g., clinical investigators) to explore and evaluate network’s data fitness for a particular use case.

Funding

This work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-04608) for development of the National Patient-Centered Clinical Research Network, known as PCORnet.

Disclaimer

The statements presented in this publication are solely the responsibility of the author(s) and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee or other participants in PCORnet.

References

Data Science Approach to Monitoring Health Status for a Diabetes Population

Amanda Ford, PhD1, Wei Lu, PhD1, Jennifer Schneider, MD, MS1, Jesse Bridgewater, PhD1, Jenna Bollyky, MD1,2

1Livongo Health, Mountain View, CA 2Stanford University School of Medicine, Stanford, CA

Introduction

Connected medical devices such as glucose meters, activity trackers, scales, and blood pressure cuffs offer healthcare providers an opportunity to monitor a patient’s health status outside of the clinic. The cellular-enabled glucose meter offered through the Livongo for Diabetes program captures real-time self-monitoring of blood glucose (SMBG) measurements and contextual data with each blood glucose (BG) check for over 50,000 people with diabetes.

Individual SMBG patterns are important for medication adjustments and patient behavior feedback. However, clinicians primarily rely on the hemoglobin A1c lab test (A1c), a measure of BG control over the prior three months, for health status monitoring, risk stratification and treatment interventions for patients with diabetes. A1c levels above 7% are associated with higher rates of diabetes complications in multiple large-scale clinical trials and normalization of A1c levels correlates with reduced complications.1 Several studies supporting a linear relationship between mean BG and A1c have been published with correlation coefficients ranging from 0.71 to 0.86. However, these incorporate SMBG frequency data ranging from two to seven times per day.2-4 In our real-world setting, however, many Livongo members perform SMBG less often than two times daily as shown in Figure 1A.

We hypothesized that a robust A1c could be estimated for individuals with sparse SMBG data by applying a novel statistical approach to support personalized member outreach in a clinically meaningful way.

Methods

Lab A1c values obtained from 559 Livongo members on the program for at least 90 days were compared to an estimated A1c (eA1c) calculated on the same day as the lab value was drawn using the ADAG study model where eA1c = [mean BG over past 90 days + 46.7]/8.7. 3 Two population restrictions based on SMBG frequency were made to improve the accuracy of the eA1c. First, the study population was restricted to members who had a minimum of 90 SMBG checks over 90 days prior to the lab measurement. Second, we implemented a power analysis to determine the minimum number of SMBG checks needed to estimate A1c for a given a set of assumptions. Because we are interested in observing meaningful changes in health status in order to risk stratify and personalize member outreach, we set a personalized minimum detectable effect size for changes in A1c based on a change that would be clinically meaningful for someone at a given A1c level (Figure 1B). For example, a member with an elevated eA1c of 12% would need to reduce their eA1c by 17% (2 points) to detect eA1c improvement with a given acceptable rate of false positives and false negatives. Across all A1c values, the minimum detectable effect is 11%. The distribution of eA1c for the entire Livongo population and population restricted subsets is shown in Figure 1C.

![Figure 1](image-url)

Figure 1. (A) SMBG checking frequency over past 90 days for the Livongo population. (B) Clinically-determined minimum detectable effect size assigned to A1c values to increase observability. (C) Distribution of estimated A1c for the Livongo population.
Using the power.t.test package in R\(^5\), we assume an effect size as shown in Figure 1B, a significance level (\(\alpha\)) of 0.2, power (1-\(\beta\)) of 0.8 and a standard deviation derived from a member’s SMBG checks over the previous 90 days, to solve for \(N\), the number of checks an individual would need to detect a change in A1c, should one occur. If a member’s actual SMBG checks in the previous 90 days equaled or exceeded \(N\), the individual was considered “observable” with a reliably estimated A1c.

We evaluated the fraction of the entire population that met the applied SMBG restrictions as well as accuracy metrics for the model when applied to each population including mean absolute deviation (MAD), mean absolute relative deviation (MARD) of estimated A1c from lab HbA1c, root mean square error (RMSE), correlation, R-squared, and error grids for lab A1c vs. estimated A1c.\(^4\)

**Results**

The fraction of members who met SMBG criteria was higher for observable members vs. members who performed SMBG \(\geq 90\) times over 90 days (46%, \(N=259\) vs 26%, \(N=147\), respectively). Metrics of accuracy for the model applied to each population are presented in Figure 2A and are broken out by A1c range in tables below each panel (Figures 2B-D). Zone A for eA1c accuracy was defined as eA1c within 10% of the lab A1c value or both lab A1c and eA1c were <6% or >10%. Zone B was defined as eA1c within 20% of the lab A1c value. eA1c values outside of the A+ B zones are generally considered erroneous. For both population restricted models, 86% of eA1c values fell within 20% of the lab values. RMSE was higher on average for observable members but values are sufficiently low, i.e. 0.5-1.0%, around the clinically important target A1c of 7%.

![Figure 2](image-url)

**Figure 2.** (A) Accuracy metrics for the study populations including mean absolute deviation (MAD), mean absolute relative deviation (MARD) of estimated A1c from lab HbA1c, root mean square error (RMSE), correlation, R-squared and % of estimated A1c values within 10% (Zone A) and 20% (Zone B) of lab A1c values. Error grids for lab A1c vs. estimated A1c for (B) all members (C) members with \(\geq 90\) SMBG checks over past 90 days and (D) observable members.

**Conclusion**

We present a clinically-relevant framework for estimating A1c from sparse SMBG data captured via cellulary-connected glucose meters in a large population of people with diabetes. Across the entire Livongo population, 25% of active members have \(\geq 90\) SMBG checks over the past 90 days, whereas 53% are observable. This approach more than doubles number of people whose health status can be observed remotely, with similar accuracy to a model with more stringent SMBG checking requirements. These findings will allow Livongo to support personalized member outreach in a clinically meaningful way.

**References**

A Model for Supplementing EHR Data with Location-based Information within an i2b2 System

Bret J. Gardner¹, Jay G. Pedersen, MA¹, James R. Campbell, MD¹, Mary E. Campbell, PhD², W. Scott Campbell, PhD, MBA¹ James C. McClay, MD, MS¹
¹University of Nebraska Medical Center, Omaha, Nebraska; ²Texas A&M University, College Station, Texas

Introduction
Data available within the electronic health record (EHR) is insufficient to provide a complete picture of an individual patient. Information detailing a patient’s social, economic, and environmental determinants of health (healthypeople.gov), current vital status, and specific details about disease are often absent. However, these data are available from external sources. Information related to neighborhood socioeconomic status (NSES) is available in U.S. census data and current vital status and detailed disease information are often recorded in government, trial, or disease registries. Socioeconomic status is a major contributing factor in readmission, morbidity, and mortality (1). With the paucity of data in the EHR related to socioeconomic status, historically researchers have used insurance type as a proxy for this measure (2). Models of socioeconomic status built from diverse U.S. census data have been used by public health researchers. To date, no systematic approach has been demonstrated to link U.S. census data to patient data within an i2b2 de-identified clinical data warehouse. We demonstrate a model to incorporate socioeconomic data from the census into i2b2.

We hypothesize location-based, extra-EHR data may be incorporated into a clinical data warehouse and subsequently de-identified. Specifically, we propose a neighborhood socioeconomic status indicator may be created from variables available within the American Community Survey (ACS) and linked to patients using geocoded addresses.

Methods
Linking geocoded addresses to the ACS: Patients were linked to neighborhood socioeconomic information within the ACS by geocoding their addresses to obtain the block group geographic identifier defined by the ACS. This geocoded information was provided for the local database by researchers at the University of Minnesota. Geocoded patient information was maintained in a database that was inaccessible to clinical researchers. Only aggregate and summary data were made available to researchers via queries using Informatics for Integrating Biology & the Bedside (i2b2 – www.i2b2.org).

ACS variables: Socioeconomic indicator variables were selected from the ACS based on a model validated by Bird et al. (1). These variables (Table 1) were extracted from the 5-year estimates from 2011-2015. We standardized the raw values from ACS to a mean of zero and a standard deviation of one.

Population: We examined the subset of patients included in the EHR residing in Nebraska and Iowa, which includes the majority of patients receiving care at our local health system.

Table 1. Description of ACS variables extracted and percent of geocoded patients linked to each variable

<table>
<thead>
<tr>
<th>ACS Variable</th>
<th>Description</th>
<th># of Patients</th>
<th>% of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>B15003</td>
<td>Percent of adults older than 25 with less than a high school education</td>
<td>1,426,786</td>
<td>100.00%</td>
</tr>
<tr>
<td>B17017</td>
<td>Percent of households with income below the poverty line</td>
<td>1,426,481</td>
<td>99.98%</td>
</tr>
<tr>
<td>B19013</td>
<td>Median household income</td>
<td>1,409,500</td>
<td>98.79%</td>
</tr>
<tr>
<td>B19057</td>
<td>Percent of households receiving public assistance</td>
<td>1,426,481</td>
<td>99.98%</td>
</tr>
<tr>
<td>B23007</td>
<td>Percent of female-headed households with children</td>
<td>1,426,263</td>
<td>99.96%</td>
</tr>
<tr>
<td>B23022</td>
<td>Percent male unemployment</td>
<td>1,426,786</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Data Warehouse: Prior to this study, the local clinical data warehouse contained information extracted from the clinical enterprise EHR. We augmented the data warehouse to include the six standardized ACS socioeconomic
variables needed to calculate the index described by Bird. We created metadata for these variables allowing them to be queried in i2b2. To demonstrate the utility of these combined data, we performed a query evaluating emergency department utilization stratified by neighborhood socioeconomic status.

**Results**

We successfully linked 1,426,786 patients with a Nebraska or Iowa address to block group information from the ACS. Table 1 documents the percentage of these patients having block group level estimates for each variable within ACS. Each variable represented information required by a validated socioeconomic index (1). In all cases, the percentage exceeded 98%.

Using these data, we demonstrated a result consistent with previous findings (3). From i2b2 (Figure 1) we identified patients with an emergency department encounter during the years 2013 to 2016, for whom we had linked neighborhood socioeconomic status variables (Figure 2). This showed usage of the emergency department to be skewed nearly 3:1 toward individuals living in neighborhoods with below average median income.

**Conclusion**

We successfully linked extra-EHR data to existing patient information within our clinical data research warehouse and used i2b2 to demonstrate a result consistent with the literature. Using Bird’s variables for neighborhood socioeconomic status, we have provided a model to incorporate ACS data in a query-able fashion for clinical research. Using this approach, we plan to identify additional socioeconomic variables and indices to increase the research capabilities of the clinical data warehouse.

**References**


Introduction

Randomized clinical trials (RCTs) are the gold standard for determining the treatment effect for a new intervention such as a medication. Within the causal inference community, the estimated treatment effect from an RCT is referred to as the “Average Treatment Effect” (ATE). This average is with respect to the clinical trial sample. However, if there is any treatment heterogeneity (e.g. the treatment has differential benefits for older vs younger people) the ATE will vary across different populations.

This is a particular concern within RCTs, because it has been well documented that RCT samples often differ from those that actually use the drug (see for example\(^1\)). The authors, compared the HgbA1C distribution for over 200 diabetes trials to diabetics in their institution’s electronic health record (EHR) system and found that patients in diabetes trials tended to have higher HgbA1C than those in the real clinical population. For this reason there has been considerable interest in estimation of “Population Average Treatment Effect” (PATE). Here, one specifies a target population, such as a clinical population seen at one’s medical center, at estimates the treatment effect in that target. This allows a health system to determine what is the expected treatment effect for their population.

Most of the work on translating a RCT result from a source trial to a target population has emanated from the epidemiology, causal inference literature. As such, most of the methods employed have focused on reweighting based approaches\(^2\). However some work has suggested that a machine learning/predictive model approach may perform better\(^3\).

In this talk we will assess the ability to translate the results from an RCT to a target population defined through our institutions EHR. To perform the translation, we utilize the Random Forests (RF) algorithm and extend the methodology of Causal Random Forests\(^4\).

Methods

Our source RCT consists of the NAVIGATOR trial, a 2X2, factorial trial for treating pre-diabetes\(^5\). We consider two of the comparisons explored by the trials: Valsartan vs Nateglinide and Valsartan versus Placebo. We used the primary outcome of incident diabetes as our outcome of interest. The study authors reported small but significant treatment effect for Valsartan in both comparisons.

We identified pre-diabetic patients into our institution’s, Duke University Medical Center (DUMC), EHR to serve as our target data source. To identify this target data source, we had to identify patients who theoretically would have been eligible for the NAVIGATOR trial. We followed the inclusion/exclusion criteria outlined in the trial protocol. Beyond meeting the clinical characteristics, we wanted to ensure that DUMC was the patient’s medical home. Therefore we also required patients to live local to our institution and to have had at least two encounters in the past two years.

To translate the RCT results to our clinical population, we adapted the Counterfactual Random Forests approach by Lu et al\(^4\). In brief, for a two arm trial, i.e. Valsartan vs Placebo, one builds a RF prediction model for diabetes among those people that received Valsartan and an RF model separately for those that received Placebo. Then, one takes the EHR samples and “runs” them through each model. This produces a predicted probability under each condition for all people, the counterfactual. For each person one can take the differences between these probabilities to estimate an individual risk difference and then average over all people to get the PATE. We used 1000 bootstrap samples to estimate standard errors.
**Results**

We identified 13,152 people who theoretically would have met the inclusion criterion for the NAVIGATOR trial, and were part of our “local” patient population. There were meaningful differences between our patient population and the trial population. Some differences included: the clinical population being younger, more African American, heavier, less a history of CVD, and greater usage of certain medication.

Table 1 provides the results for estimating the Risk Difference directly in NAVIGATOR (Empirical), applying Causal RF back to the NAVIGATOR sample (RF in NAV) and then translating the trial results to Duke EHR (RF in EHR). One can consider the ability for the second analysis to replicate the first as an internal validation of the methodology. The results suggest that the Causal RF methodology was able to replicate the empirical estimates. More-so, we see that while there is a slight treatment effect in the Valsartan vs Nateglinide comparison within NAVIGATOR, this effect goes away within the Duke EHR. Conversely, while the Valsartan vs Placebo effect is not perfectly replicated (the 95% CI crosses 1) we note that the difference between the two estimates is not significantly different, suggesting that there may be some treatment benefit to Valsartan vs Placebo.

<table>
<thead>
<tr>
<th>Risk Difference</th>
<th>Val vs Nat</th>
<th>Val vs Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical in Nav</td>
<td>-0.04 (-0.06, -0.01)</td>
<td>-0.05 (-0.08, -0.02)</td>
</tr>
<tr>
<td>RF in Nav</td>
<td>-0.04 (-0.06, -0.02)</td>
<td>-0.05 (-0.06, -0.02)</td>
</tr>
<tr>
<td>RF in EHR</td>
<td>0.01 (-0.02, 0.05)</td>
<td>-0.02 (-0.06, 0.02)</td>
</tr>
</tbody>
</table>

**Table 1**: Estimated Risk Difference (and 95% CI) in NAVIGATOR, after translation to the EHR.

**Conclusion**

This work illustrates how one can combine methodology from both the machine learning and causal inference communities to translate results from a clinical trial to a target population. The advantage of the RF approach is that it is able to model more complex heterogeneity (i.e. higher level interactions). Most trials report a Supplemental Table 1, assessing treatment heterogeneity. NAVIGATOR did this and reported no interactions. However these are only first order interactions on dichotomized variables. The use of RF allows one to model subtler effects.

An ML approach has an advantage over the epidemiological (weighting) approach because one can apply the generated model to any target population - even a single person. Conversely, the weighting approach needs to be adapted to each target. While other ML methods can be used, the bootstrapping of RF provides some methodological advantages. However the ML approach does require a larger RCT sample size to properly estimate the model.

As health systems obtain the ability to better understand their patient populations, the desire to understand how treatments will impact their patients will increase. These decisions may potentially become more salient as health systems act more and more as defacto insurance companies, creating drug formularies for their patients. Such an ability to translate trial results to a patient population can help inform which drugs to prioritize.

**References**

Secure Reusable Linkage Between a EHR Patient Portal and REDCap for Population-based Survey Research

Melissa A. Habrat, BS, Lucas D. Moriera, BA, John Clark, BS, Candace Jaruzel, PhD, MSN, CRNA, James C Oates MD, Jihad S Obeid, MD, Leslie A Lenert, MD, MS

Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC USA

Abstract

Conducting survey research in medical populations is often difficult. We describe an approach for automation of the process using population health tools in Epic and a REDCap modifications that create a dedicated personal survey queue for patients. Results show that high-response survey rates (50%) are possible even without incentives but that further work is needed to expand institutional research permissions registries so that they are more representative of patients as a whole.

Introduction

Administration of research surveys using patient portals poses complexities. Research surveys normally are not included in the legal medical record and therefore the use of a secondary application, such as REDCap, is a preferable alternative to storage in a “research only” portion of the record. Direct communications to patients can lead to issues with lack of authentication of participants and participants potentially not recognizing the electronic sender and ignoring the survey request. The resulting lack of certainty about the respondent population and low response rate can negatively impact the power of the research study and/or may result in study failure. To mitigate those issues at the Medical University of South Carolina (MUSC), we created a set of systems would allow research teams to send invitations to participate REDCap surveys through the secure Electronic Health Record (EHR) patient portal (MyChart). This was accomplished by utilizing a participant-specific unique Uniform Resource Locator (URL) linked to a patient specific survey queue in REDCap.

Methods and Evaluation

The design of the system is shown in Figure 1. Epic patient registry tools were used to create and store a patient specific URL linked to REDCap. Specific studies create a subregistry based on clinical criterial in Epic. Patients in the sub registry receive a MyChart notification message. When patients sign into MyChart, identifying themselves, they are taken to an explanatory screen for the survey with the custom link to the patient’s REDCap queue. Clicking on this link takes the patient to his or her survey queue in REDCap. This is resolved by REDCap based upon a coded identity in the URL for the survey. Patients then complete a REDCap survey as usual. A coded identifier allows re-linkage with clinical data if required. Alternatively, data can be anonymized and exported by an honest broker.

Figure 1. Workflow for administration of surveys using Epic’s population health tools, MyChart and REDCap
We evaluated this workflow in a simple survey conducted with the Department of Anesthesiology. They created a survey in REDCap to explore potential preoperative anxieties that patients who had recently had surgery had experienced. They developed criteria for identifying patients in Epic and a sub-registry for their task, and a REDCap survey. Surveys were emailed to patients who were older than 18 years of age who underwent moderate risk surgeries and required general anesthesia with endotracheal intubation in the past month. They then linked their patient criteria to the MUSC “I am changing what is possible” research volunteer registry (1) to patients who, in addition to meeting clinical criteria for recent surgery, had provided permission for direct patient contact for research using MyChart or in clinic. These patients were sent a invitation to participate in this survey research via MyChart, followed by two reminders. Eligibility and response rates are described.

Results

Results from the survey are show in Figure 2 in the form of a CONSULT diagram. About 608 patients met criteria for recent survey and were added to the study’s sub-registry. Of these, 318 (50%) had MyChart accounts. Only 80 of the patients with MyChart accounts had enrolled in the “I’m changing what is possible” registry (25%). Of these, 40 responded after three contact attempts (50%). Thirty-eight of the respondents (95%) had usable data.

Discussion

This study demonstrates how Epic population health functionalities can be adapted for research. Through the use of a patient specific URL for all registered patients stored in an Epic Healthy Planet registry with patient-specific survey queues in our REDCap application, we were able to securely link Epic’s clinical data and patient outreach and authentication capabilities with used functionality with REDCap to administer online research questionnaires. The coded identifier allows allows auto login to REDCap, eliminating the need for usernames, passwords, or other identifiers.

MyChart has limited survey capabilities compared to other dedicated survey research platforms. This approach allows users to rapidly extend those capabilities using in REDCap. Currently there are limitations on how to get data base into Epic from REDCap surveys. In the future, application programming interfaces such as the Fast Healthcare Interoperability Resource (FHIR) may make this easier. At present, one easy and reliable way to do this, albeit in a non-real-time fashion, is Epic’s DataLink protocol.

The results of this study show that the generalizability of this approach to survey research is limited by the size of patient research volunteer registries. When this work was conducted, MUSC’s registry had about 30,000 participants, out of about a 350K clinical population. Selection bias in persons who opt-in the registry may result in potential bias in survey results. (cite to registry paper).

More optimistically, the response rate among patients contacted who were in the registry was 50%. This suggests that meaningful work can be done with research registries if the registries themselves are of adequate size and are representative. Supplementing MyChart messages with telephone contacts or other tools might improve recruitment rates. This study did not offer incentives to patients and use of a cash incentive or lottery incentive might have dramatically raised response rates.

References


Acknowledgments: This work was supported in part by the South Carolina Clinical & Translational Research (SCTR) Institute, with an academic home at the Medical University of South Carolina, NIH/NCATS Grant number UL1 TR001450
Administration of Patient Reported Outcomes (PRO) Questionnaires in Clinical Settings Using Provider Workstations

Melissa A. Habrat, BS, Lucas D. Moriera, BA., James C Oates MD, Jihad S Obeid, MD, Leslie A Lenert, MD, MS

Biomedical Informatics Center, Medical University of South Carolina, Charleston, South Carolina

Abstract

Capturing patient reported outcome measures in clinical settings is often difficult because of the requirements for specially purposed hardware and software tools. We describe our experience adapting Epic’s MyChart Patient questionnaire tools for use in clinic by patients on exam room computers after rooming of patients by ancillary providers. The process was well accepted by both patients and providers and allows MyChart tools to be used both for PRO capture and implementation of study eligibility criteria.

Introduction

Collection of patient administered e-questionnaires in clinical settings is often difficult because of simple logistical barriers. If a separate computer system, such as a laptop or tablet is used, clinics need to maintain an inventory of these devices and face risks of loss. Costs of purchase can be a substantial barrier to implementation, particularly for research questionnaires. Electronic health record system developers and other vendors have developed kiosks for this purpose, often positioned in waiting rooms. However, these devices pose similar challenges to use of tablets in terms of cost of devices and, in some cases, costs for data capture on a per patient basis. If free-standing computer systems are used, authentication of patients can become an issue. Integration of results back into the EHR is also often complex. In this presentation, we describe our efforts to adapt the Epic MyChart questionnaire administration system in the Epic EHR for use by patients of the exam room computer as a kiosk during the period between rooming of the patient and by ancillary personnel and the provider visit.

Methods and Evaluation

Our questionnaire process is based on the time between the rooming of the patient and the provider’s visit with the patient as being dedicated to patients answering questionnaires. This time, in our institution, is opportune. There is increasing emphasis on quickly moving patients from waiting rooms to the exam room. The examination room is a relatively private place to administer questionnaires (much more so than a public waiting room.) The computer in the exam room is not utilized during this period. Each patient in our Epic implementation has an active queue of questionnaires. When the patient questionnaire option is selected from the menu, this brings up, directly, the MyChart Questionnaire window and hides the provider window and locks other functionality of Epic. After completion of the questionnaires in the queue, the screen is entirely locked,
except for the provider log in box, enforcing good security practices. At any point in the questionnaire, if the provider needs to see the patient, the process can be interrupted and the screen returned to provider control. The approach has the additional advantage of recording data directly into the Epic database, where results can be easily used to trigger alerts for providers and other data collection process. The process of conversion the provider EHR to a patient kiosk-like device is shown in Figure 1.

We have tested this workflow extensively in a number of settings: use of this method to capture research and surplus tissue use permissions in clinic settings (1), use of the method to patients to self-administer an eligibility screening questionnaires to screen patients for eligibility for a pragmatic trial of physical therapy for back pain and for use (with optional tablets) to administer PROs (HIV-36) in an infectious disease specialty clinic. Questionnaires can be set to appear in a patient’s queue based on time elapsed since the last questionnaire, clinical factors and responses to previous questionnaires. The system can be used with questionnaires completed on tablets when clinic regulations require this. In MUSC’s ophthalmology and HIV clinics, policies either forbid the use of exam room computers by patients (HIV) or make it impractical (in ophthalmology clinics, patients cannot sit on providers’ rolling stools.)

Results

The research preference questionnaire is in active use in approximately 49 clinics across MUSC. On a monthly basis, the number of questionnaires collected is about half that through MyChart (645). The workflow is well accepted in clinics and the usual reason for non-usage is registration of the patient for the MyChart portal.

The same workflow is in place in all general medicine clinics to evaluate patients with back pain as a chief complaint for the PCORI sponsored TARGET pragmatic trial. A chief complaint of back pain triggers the addition of the questionnaire to the patients questionnaire queue. After patients complete a screening questionnaire that determines trial eligibility, the provider administers a second questionnaire to patients on symptoms and quality of life. More than 1400 patients have used the screening questionnaire. In HIV clinics, we are piloting administration of the HIV-36 in one providers patient group with plans to expand to two more.

Discussion

Capture of patient reported outcomes is becoming a cottage industry in healthcare. The ability to rapidly add patient reported data within care systems without the purchase of new equipment can greatly accelerate the conduct of clinical and translational research. The approach takes advantage of a neglected period in clinical care (the time between rooming and the provider visit) that offers privacy, time for contemplation, and the potential for staff assistance when needed. Our results show that use of the provider’s workstation ask a kiosk like device for patients is practical in many instances and well accepted in clinical environments. It can be used both for population based studies and to screen patients for pragmatic trials and other randomized studies. While this might not be practical in every clinic, it can be combined with use of other devices such as iPads to support patient input when necessary. Creation of a survey queue for patients and integration with population health tools in the EHR make the approach flexible and easily integrated with other clinical objectives.

References


Acknowledgments: This work was supported in part by the South Carolina Clinical & Translational Research (SCTR) Institute, with an academic home at the Medical University of South Carolina, NIH/NCATS Grant number UL1 TR00145
Pediatric Emergency Assistance for Newborns Using Telehealth (PEANUT): Design and implementation of a neonatal telemedicine program

Sarah C. Haynes, MPH1; James P. Marcin, MD, MPH1; Kristin Hoffman, MD1; Mark Underwood, MD, MPH1; Barry Hill, MPH1; Namita Kansal, BA1; Madan Dharmar, MBBS, PhD1.

1University of California Davis Health, Sacramento, California.

Introduction: Access to neonatal subspecialty and intensive care services is associated with lower mortality for high-risk infants. While rural hospitals have the capacity to fulfill the basic healthcare needs of newborns, they have extremely limited access to pediatric subspecialty care, which may contribute to higher mortality rates. Telemedicine has the ability to reduce unnecessary transfers and decrease delays in diagnosis and intervention for critically ill infants by providing remote access to neonatologist and other pediatric subspecialists. Telemedicine technology allows rural providers to access by allowing specialists to provide a rural provider with a plan that allows the newborn to be managed locally. Additionally, telemedicine can also be used to improve screening for critical congenital heart disease (CCHD). California state law mandates pulse oximetry screening to reduce the mortality and morbidity associated with CCHD through early detection and intervention. For infants who fail the pulse oximetry test, Centers for Disease Control and Prevention (CDC) recommends obtaining a diagnostic echocardiogram, which is often unavailable at rural hospitals. The use of telemedicine for echocardiography can prevent unnecessary transfers to tertiary care facilities to conduct these tests. Although the benefit of telecardiology services has been widely demonstrated, there is a scarcity of literature on the impact of telemedicine for neonatology and other pediatric subspecialties.

The Pediatric Emergency Assistance to Newborns Using Telehealth (PEANUT) program at the University of California Davis Children’s Hospital (UCDCH) was started with grant funding from the Office for the Advancement of Telehealth- Health Resources and Services Administration (OAT-HRSA) with the goal of improving healthcare and reducing disparities among newborns delivered in rural hospital nurseries.

Methods: The Pediatric Emergency Assistance to Newborns Using Telehealth (PEANUT) is a comprehensive telehealth program that provides continuous access to neonatologists and other pediatric subspecialists to a network of six rural hospital nurseries. PEANUT supports care for newborns in rural hospitals by providing access to neonatologists and other pediatric specialists through telemedicine and phone consults. These consults may include diagnostic and clinical support as well as interpretation of screening tests that may indicate a transfer, including pulse oximetry and echocardiograms. The program also seeks to improve the ability of rural providers to better treat newborns in their own communities by providing live monthly training and ongoing education to rural providers, including physicians, nurses, and technicians. Consultations may occur through telephone or telemedicine.

The PEANUT program (Figure 1) is integrated into the existing UC Davis Transfer Center enabling referring physicians from rural hospitals to call the program if a newborn is in need of a neonatology or pediatric sub-specialty support. The transfer center connects the referring provider to the neonatologist for the initial assessment to determine if the newborns needs to be transported to a neonatal intensive care unit (NICU) for care. If the newborns needs to be transported the neonatologist dispatches the transport team and continues the consultation on telemedicine. If the newborn does not need a transfer to NICU, the neonatologist continues the consultation either by phone or telemedicine based on the individual needs of that newborn. Telemedicine consultation is conducted using a dedicated high-resolution videoconferencing unit.
These units have embedded encryption to ensure patient privacy and safety. Each telemedicine consultation involves a live, interactive audiovisual communication between pediatric specialists and referring providers with the newborn and the parent or guardian when available. Consulting neonatologists engage other pediatric subspecialists in the same telemedicine consultation on an as needed basis. PEANUT also serves to provide echocardiography through telemedicine for newborns that do not pass the pulse oximetry screening for CCHD or who have suspected CCHD. The neonatal echocardiogram is conducted at the remote hospital and sent to UCDCH pediatric cardiologist by store and forward telemedicine to Syngo Image Server located at UCDCH. The pediatric cardiologist reviews the echocardiogram and advises a transfer to further care if the newborn has a CCHD. At UCD we created a separate encounter within EPIC electronic health record to document all telemedicine consultations. These encounters are processed through the billing system like any other specialist consultations for newborn admitted at UCD. These telemedicine consultation record is made available to the referring remote physician through EPIC’s Physician Connect.

Figure 1. Framework for the PEANUT program.

Results: Since the launch of the program, PEANUT has provided specialist consultations to a total of 302 newborns at six rural hospitals in Northern California. 68 of these consultations were telemedicine consultation and 234 were telephone consultations. PEANUT also provided pediatric cardiology consultations for echocardiogram for 211 newborns who either failed pulse oximetry screening for critical congenital heart defects or were suspected to have a congenital heart defect. Of the 302 newborns who received a telemedicine or phone consultation, 227 (75.2%) were transferred to a tertiary care facility. Transfer rates varied by site, ranging from 53.1% to 88.3%. Telemedicine consults were less likely to result in a transfer than phone consults (63.2% vs 78.6%, p=0.010).

Discussion: PEANUT may be effective in reducing unnecessary transfers to tertiary facilities. PEANUT provides timely access to pediatric subspecialists for care in the patient’s own community; it also provides education and training to healthcare providers in the rural hospital, which increase the level or competency of care that can be provided locally. PEANUT demonstrates a new model of care that can be adopted as a clinical service line for other children’s hospitals in a competitive healthcare market.
Architecture of a visual interactive analysis tool for filtering and summarizing large data sets coded with hierarchical terminologies (VIADS)

Xia Jing, MD, PhD1, Matthew Emerson2, Drew Gunderson, BS2, James J. Cimino, MD3, Chang Liu, PhD2, Jay Shubrook, DO4, Brian O. Phillips, MA1

1Department of Social and Public Health, College of Health Sciences and Professions; 2Russ College of Engineering and Technology; 3Informatics Institute, School of Medicine, University of Alabama, Birmingham, Alabama; 4College of Osteopathic Medicine, Touro University, Vallejo, California

Introduction

Enormous volumes of data, coded through hierarchical terminologies (such as the International Classification of Diseases 9th revision-Clinical Modification [ICD9-CM], ICD10-CM, the Systematized Nomenclature of Medicine-Clinical Terms [SNOMED-CT], Logical Observation Identifiers Names and Codes [LOINC], RxNorm, and the Gene Ontology [GO]), are generated continuously within electronic health record systems and in literature databases (such as Medical Subject Headings [MeSH] in PubMed) and other sources. The large size of these data sets and complex terminologies’ structure make them difficult for clinicians, clinical researchers, or administrators to understand. Further processing and analysis of these data sets are needed to make them manageable and comprehensible.

Our group has published the core algorithms to filter, summarize, and compare large data sets coded by hierarchical terminologies1, methods to set thresholds and demonstrated case studies2, and preliminary results of the comparison of two solutions to develop an online tool for the algorithms3. We are developing a Web tool, i.e., VIADS (https://www.viads.info/) to make the algorithms publicly accessible for free. The tool can be utilized by clinical researchers, healthcare administrators, clinicians, and biomedical trainees to conduct secondary data analysis, filtering, visualization, hypotheses generation, and validation, including exploring new patterns, new facts, or relationships by looking at the aggregated effects within the data sets. The summarizing, filtering, and comparing capabilities provided by VIADS cannot be accomplished by any single existing tool. This manuscript introduces the architecture design of VIADS (Figure 1). Table 1 provides a summary of the algorithms that we are implementing and their usage examples. Figure 2 illustrates the work flow of VIADS’ analytic engine.

![Architecture of VIADS](image)

The acceptable data sets of VIADS have to meet two criteria: data need be coded using a hierarchical coding system (i.e., ICD9-CM), and frequencies need to be available for each code. The output files of VIADS are filtered graphs (PNG and SVG) and the corresponding data file (CSV). A parent-child table of the terminology hierarchies is necessary to generate graphs and to conduct further calculating. Currently, we have implemented the visualization module successfully. We also have implemented all algorithms in Table 1 in the analytic module. We are in the process of connecting different modules and migrating the tool from our local server to a Web server now. The main tools and
development environments that we have utilized include Django, Vis.js, JQuery, Javascript, Chart.js, Unittest, R, and MySQL. The visualization module provides alphabetic nodes sorting, various spacing options between hierarchical levels, automatic resizing of the graphs, a pop-up box with extended information about the node when hovering over it, and color scales to reflect exact data behind the nodes and edges. The visualization module can fit 130 nodes easily and efficiently in an 11-inch laptop screen. We will post a data preparation document including SQL queries to guide users to prepare ICD-9, ICD-10 datasets from their source databases.

Table 1. Algorithms to be implemented in VIADS and examples of their usage

<table>
<thead>
<tr>
<th>Filter</th>
<th>Definition</th>
<th>Usage example</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC (node counts)</td>
<td>NC = usage frequency of a node (ICD code or MeSH term)</td>
<td>Displaying a summary view: the most frequently used MeSH terms and their ancestors in 2011</td>
</tr>
<tr>
<td>CC (class counts)</td>
<td>CCnode = NCdescendant1 + NCdescendant2 + NCdescendant 3…</td>
<td>Displaying a summary view: the most frequently used ICD9-CM codes in 2011 in a selected institution</td>
</tr>
<tr>
<td>Ratio</td>
<td>Ratio = CCchild node/CCparent node</td>
<td>Identifying the largest MeSH contributors to upper-level MeSH terms and their ancestors in 2011</td>
</tr>
<tr>
<td>Systematic comparison</td>
<td>CCnode1 vs CCnode1; CCnode2 vs CCnode2; CCnode3 vs CCnode3; … …</td>
<td>Displaying a comparison view: the most significant different ICD9-CM codes between pioglitazone(data set1) and rosiglitazone (data set2) groups after systematical comparison</td>
</tr>
<tr>
<td>Combination</td>
<td>NC + Ratio</td>
<td>Displaying a summary view: the most frequently used MeSH terms and the largest MeSH contributors and their ancestors in 2011</td>
</tr>
<tr>
<td></td>
<td>CC + Ratio</td>
<td>Displaying a summary view: the most frequently used ICD9-CM codes and the largest ICD9-CM contributors and their ancestors in 2011</td>
</tr>
</tbody>
</table>

We plan to make VIADS accessible to both guest users and registered users for free for non-commercial usage. VIADS can be utilized to generate a new dimensional perspective about data sets, which, in turn, can be utilized to facilitate more informed administrative decisions (e.g., to allocate resources), research decisions (e.g., to validate or deny hypotheses), or clinical decisions (e.g., to select similar medications based on analysis of aggregated data sets).

Figure 2. VIADS analytic engine work flow

Conclusion

The rapid adoption of electronic health records in both office-based practices and hospitals has led to an increasing number of available coded data sets. These coded data are becoming increasingly common not only on the administrative side (e.g., for billing purposes) but also on the clinical side (e.g., to generate a problem list). The development of such a publicly accessible tool will help users to achieve a summary view, secondary analysis, and visualization of their data sets with minimal technical effort. The utility and usability of the tool need further study.

References

Beyond Cohorts: Multi-institution Datasets in the Cloud

Jeffrey G. Klann, PhD1,2,3, Stanley Boykin4, Margaret Vella1, Marc Natter, MD1,4, Sebastian Schneeweiss, MD, ScD2,3, Douglas Macfadden1, Sarah Weiler, PhD1, Kenneth D. Mandl, MD, MPH1,5, Shawn N. Murphy, MD, PhD1,2,3

1Harvard Medical School; 2Partners Healthcare; 3Massachusetts General Hospital; 4Brigham and Women’s Hospital; 5Computational Health Informatics Program, Boston Children’s Hospital. – All in Boston, MA

Introduction
ARCH, the Accessible Research Commons for Health, is a Clinical Data Research Network (CDRN) in the Patient Centered Outcomes Research Institutes’ PCORnet network. (ARCH was formerly known as SCILHS, the Scalable Collaborative Infrastructure for a Learning Health System.) It has 10 million patient lives at 10 sites across six states. [1,2] For over two years, the network has enabled live preparatory-to-research queries using the Shared Health Research Informatics Network SHRINE platform, which links together Informatics for Integrating Biology and the Bedside (i2b2) data warehouses across institutions. This platform enables researchers from participating institutions to execute complex cohort finding queries on the entire network, live, and without human authorization for individual queries. [3] (Queries are audited in retrospect.) The SHRINE platform allows researchers to get a very specific sense of what data is available in the network.

ARCH has also developed tools to facilitate patient recruitment and surveying at individual institutions after a cohort has been found, collaborating physicians have been secured, and protocols have been approved by IRBs. However, these two use cases are at opposite ends of the research pipeline, and there is a large gap in the middle, in which more detailed analysis of the retrospective data from the medical record is necessary. Until now, this has been done by developing relationships with collaborators at each institution, who have access to extract the data and analyze it locally. This, however, is a time-consuming and expensive process that, while reasonable for large funded projects, is nearly impossible for preliminary research being done to support detailed feasibility analysis, usually prior to grant submission. For this reason, ARCH has developed a methodology, both technical and regulatory, to enable researchers to access limited data sets from network institutions without recruiting external collaborators and developing subcontracts. This is the first ARCH use-case in which patient data leaves individual institutions, so protecting privacy and security in the design was paramount.

Method
ARCH has developed a research portal that allows researchers in the network to request patient-level data from each site, accessed in aggregate from a secure Amazon cloud instance covered by the HIPAA Business Associate Agreement between Harvard Medical School (HMS) and Amazon.

Study data is aggregated into a HIPAA Limited Data Set (deidentified except that dates are included) through five steps, described in the following text and shown graphically in Figure 1.

1. A researcher logs into the self-service portal, develops a SHRINE query and flags it as a detailed data request.
2. Researchers select exactly what patient-level data they need through a plug-in that connects to the ARCH terminology system. The tool does not allow any data elements comprising the 18 HIPPA identifiers.
3. Once the detailed data request has been submitted, data stewards (who are employed at each site) log into their local i2b2 client and execute a second plug-in to retrieve flagged queries and extract corresponding data sets.
4. Each data steward uses a third ARCH tool to upload the data to the secure Amazon cloud instance.
5. Data are available to authorized users through an ARCH data-analysis interface during the time of the study.

Figure 1: The technical method for sharing multi-site data.
The regulatory and security processes that we have implemented to enable this technical architecture include:

1. A master IRB protocol and Data Use Agreement will be executed at every participating site. This will allow sites to upload data for approved studies to the ARCH research portal.
2. For every data set, a researcher signs a project-specific data use agreement.
3. The data is only made available through a remote desktop interface, so no data ever gets transferred to a client machine. All data download functions are disabled.

Results
At this time: ARCH is conducting a pilot implementation at Partners’ Healthcare and Boston Children’s Hospital; a beta version of the ARCH research portal has been deployed and continues to be updated; ARCH is refining the process for extracting data at our pilot sites; and, the central data-access interface is expected to be fully operational for a demo in March 2018. A screenshot of a beta version is shown in Figure 2.

Additionally, other ARCH institutions have signed the SMART IRB reliance agreement, to streamline regulatory approvals. Nine sites are live in the SHRINE platform and researchers are actively able to create accounts and run cohort-finding queries on either their local or our cloud hosted SHRINE client.

Discussion
ARCH is developing a network designed around mutual trust, and this enables agreements such as these to be enacted at our many sites with very diverse privacy regulations. The fact that these are vetted by so many diverse institutions ensures that the methods utilized will be relevant into the future. Other large research networks that have requested individual data sets have thus far done so without vetting their approach or regulatory process with their partner institutions, and this has created discord and turmoil which is eroding trust.

Figure 2. Our implementation of steps 2 and 3 above. The ARCH data selection plugin and data request handler.

Conclusion
The ARCH community-design network data-sharing methodology supports the important aspects of preparatory to research data analysis that follows cohort discovery. It is immediately implementable in SHRINE networks and methods and lessons learned can be utilized in any distributed clinical data network.

References


Abstract

Sepsis mortality is heavily influenced by the quality of care in hospitals. Comparing risk standardized mortality rate (RSMR) of sepsis patients in different states in the US has potentially important clinical and policy implications. In the current study, we aimed to compare national sepsis RSMR using an interactive web-based dashboard. The web-based dashboard (https://sepsismap.shinyapps.io/index2/) is a cross platform and publicly available to anyone with interest in sepsis outcome, health inequality, and administration of state/federal health care. After extrapolation to the national level, approximately 35 million hospitalizations were analyzed for sepsis mortality each year. Our work has the potential to support health care transparency, information diffusion, health decision-making, and the formulation of new public policies.

Introduction

The Open Government Program, created under President Obama, creates new opportunities to monitor, or to compare the performance of patient management by hospitals across the US. Recently, the Centers for Medicare & Medicaid Services (CMS) rolled out a website to compare the risk standardized mortality rates (RSMR) of acute myocardial infarction, stroke, and heart failure of individual hospital. RSMR allows comparison between hospitals as it takes into account not only patient risk factors, but also the hospital-specific mortality rate. However, the CMS did not investigate RSMR for sepsis, or use RSMR to compare inequality of sepsis care across different US states. Thus, we aimed to create a web-based dashboard for visualizing and analyzing the national annual RSMR.

Materials and Methods

This study is conducted using 2004–2011 data from the Nationwide Inpatient Sample (NIS), the largest all-payer inpatient care database in the US. We developed a cross-platform, web-accessible dashboard in six steps as depicted in Figure 1.
**Figure 1.** Workflow of Interactive Web-based Dashboard for Comparing Statewide Risk Standardized Mortality Rates of Sepsis

**Web-based dashboard Overview**

Our web-based dashboard provides a friendly environment for users to interact and visualize the 8 years of sepsis mortality data according to their needs. Data is summarized into four dimensions on the dashboard, and are switchable by clicking: Sepsis Identification Criteria; Sepsis Mortality Predictors; RSMR Map; RSMR Trend (Figure 2). In order to give the user a better experience, we implemented six interactive tools: year selection, automatic year play, map zoom, copy or print data, ranking data by name or value, and data search.

To allow users to quickly visualize the national difference in sepsis RSMR, we drew a heat map to compare the sepsis RSMR in different states. The intensity of red color was used to indicate the range of RSMR in different states, and the highest observed RSMR (40) was set as the limit. Thus, a state with a higher RSMR was colored a darker red. We colored states where data was not available as white.

**Figure 2.** Homepage of our web-dashboard, showing the “Sepsis Identification Criteria” dimension of our web dashboard, where users can compare the national sepsis RSMR for two sepsis identification criteria. We built six tools to enhance user experiences: 1. Year selection; 2. automatic year play; 3. map zoom; 4. copy or print data; 5. ranking data by name or value; 6. full text search.

**Conclusion**

This web-based dashboard allows anyone to visualize the substantial variation in RSMR across the whole US. Our work has the potential to support health care transparency, information diffusion, health decision-making, and the formulation of new public policies.

**References**

Using Misspellings in Word Vectors to Boost Information Retrieval Recall

Reed McEwan, MSSE¹, Genevieve B. Melton, MD, PhD¹, Serguei V.S. Pakhomov, PhD¹
¹Institute for Health Informatics; University of Minnesota, Minneapolis, Minnesota

Introduction

Misspellings can negatively affect the quality of information retrieval (IR) from general English¹ and from text contained in Electronic Health Records (EHR)²,³. These non-word variants are excluded from search results not only because their strings do not exactly match the query. Even stemmed versions of most misspellings do not match the stemmed or un-stemmed input. For example, searching for the word ‘serotonin’ excludes documents containing only a non-word variant (e.g., serotonin). Spelling errors in clinical documents can be attributed to the complexity of medical terminology and time constraints faced by providers, but also other non-words (e.g., abbreviations) not found in standard dictionaries or terminologies⁴. Because misspellings of a medical term (or any word) are used in similar contexts as the correct spelling, distributional semantic models can expose non-word errors. The problem is that misspellings are often buried in the co-occurrence vector due to their lower usage frequency relative to the correct spelling. We present a preliminary evaluation of efforts to improve information retrieval recall using a query expansion technique based on non-word errors contained in word embedding vectors derived from a large set of clinical notes from an EHR. Re-ranking cosine distance using edit distance is the key to finding these misspellings.

Methods

EHR clinical notes for the years 2010-2014 were used to create a word2vec⁵ model representing a 65 million document sub-corpus of the total 130 million clinical notes in our institution’s clinical data repository. The executable distance program distributed with the original word2vec software was modified to create a new program that we called spellgen that implements the Distributional Semantic Misspelling (DSM) suggestion approach. The DSM approach consists of three features added to word2vec’s distance executable to create spellgen⁶. The first modification was the calculation of Damerau-Levenshtein edit distance ($D_E$) between the input term and each word embedded in the top 1000 elements of the model’s vector for that term. The second modification was the calculation of a score, $s$, by taking the log of the exponentiated word2vec cosine distance ($D_C$) for each element normalizing by the square of its edit distance, $s = \log \frac{10^{0.5D_C}}{D_E}$, yielding values of $s \leq 1$. The third modification was the addition of data structures and a sorting algorithm for reverse sorting the vector by score. Commonly misspelled medical terms were input into spellgen to generate lists of likely misspellings. Figure 1 is a truncated screenshot of the spellgen output for the input ‘serotonin’.

Figure 1. spellgen output for the term ‘serotonin’ illustrating a skewed frequency.

A threshold for $s \equiv -0.4$ was empirically determined to be the lower limit of similarly spelled words reliably mapping to misspellings of the input term. NLP-PIER⁷, a self-service clinical documents search engine, was used to search a snowball-analyzed and indexed corpus of all 130 million patient-associated notes representing approximately 2.2 million patients. For each of the commonly misspelled terms two queries were issued: a base query consisting of a single word term and an expansion query designed to return notes containing only misspelled variants of the base query. To achieve the second set of results spellgen suggestions with $s > 0.4$ were scrubbed of words whose stems were the same as the stemmed base query term or equivalent to the stem of different, correctly spelled word. Paired queries were of the form:

| Base query:  | serotonin |
| Expansion query: | (serotonin OR serotonin OR serotonin OR serotonin OR serotonin OR serotonin OR serotonin) NOT serotonin |

Query results were manually inspected to confirm keyword matches in expansion query results were used in contexts consistent with that of its paired base query. Distinct note counts were recorded as the count of the hits for each pair of queries such that the incremental note recall is simply the expansion note count divided by the base note count. Because patients can have multiple notes, a set of distinct patient identifiers was persisted to a relational
database as a set of rows associated with each query. Incremental patient recall was computed using SQL by finding the count of expansion query patients not in the set of base query patients. Recall performance was not statistically measured for expansion queries due to a lack of an annotated set of notes.

Results

Table 1. Summary of comparing expanded vs. unexpanded query results for 18 terms.

<table>
<thead>
<tr>
<th>Term</th>
<th>Single Word</th>
<th>Expansion Terms</th>
<th>Base Query Count</th>
<th>Expansion Query Count</th>
<th>Recall Increase (%)</th>
<th>Base Query Count</th>
<th>Expansion Query Count</th>
<th>Recall Increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lidocaine</td>
<td>28</td>
<td>1741</td>
<td>826</td>
<td>0.25</td>
<td>3231358</td>
<td>3671</td>
<td>0.66</td>
<td>3231358</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>30</td>
<td>124172</td>
<td>75360</td>
<td>0.36</td>
<td>1863560</td>
<td>2256</td>
<td>0.48</td>
<td>1863560</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>15</td>
<td>67192</td>
<td>3941</td>
<td>0.48</td>
<td>1340921</td>
<td>17401</td>
<td>0.53</td>
<td>1340921</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>24</td>
<td>748260</td>
<td>1146</td>
<td>0.23</td>
<td>433758</td>
<td>14792</td>
<td>0.34</td>
<td>433758</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>10</td>
<td>28961</td>
<td>5949</td>
<td>0.36</td>
<td>1946260</td>
<td>4321358</td>
<td>0.66</td>
<td>1946260</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>23</td>
<td>122113</td>
<td>878</td>
<td>0.47</td>
<td>932737</td>
<td>1065</td>
<td>0.64</td>
<td>932737</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>25</td>
<td>441389</td>
<td>2326</td>
<td>0.49</td>
<td>1743351</td>
<td>13822</td>
<td>0.79</td>
<td>1743351</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>19</td>
<td>118898</td>
<td>998</td>
<td>0.34</td>
<td>1946260</td>
<td>4321358</td>
<td>0.66</td>
<td>1946260</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>2</td>
<td>5661</td>
<td>46</td>
<td>0.81</td>
<td>13623</td>
<td>92</td>
<td>0.68</td>
<td>13623</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>6</td>
<td>1832</td>
<td>18</td>
<td>0.87</td>
<td>20358</td>
<td>225</td>
<td>1.12</td>
<td>20358</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>11</td>
<td>107098</td>
<td>1004</td>
<td>0.53</td>
<td>919697</td>
<td>7239</td>
<td>1.29</td>
<td>919697</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>17</td>
<td>39338</td>
<td>401</td>
<td>1.30</td>
<td>2371294</td>
<td>10796</td>
<td>2.82</td>
<td>2371294</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>15</td>
<td>78765</td>
<td>1065</td>
<td>1.36</td>
<td>206442</td>
<td>21289</td>
<td>2.96</td>
<td>206442</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>22</td>
<td>184920</td>
<td>2002</td>
<td>1.28</td>
<td>49634</td>
<td>3441</td>
<td>2.96</td>
<td>49634</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>10</td>
<td>8208</td>
<td>237</td>
<td>2.87</td>
<td>8634</td>
<td>1630</td>
<td>3.51</td>
<td>8634</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>8</td>
<td>56290</td>
<td>2097</td>
<td>4.17</td>
<td>41581</td>
<td>11746</td>
<td>2.82</td>
<td>41581</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>21</td>
<td>124172</td>
<td>75360</td>
<td>5.94</td>
<td>60318</td>
<td>75360</td>
<td>12.49</td>
<td>60318</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>3</td>
<td>2007</td>
<td>10</td>
<td>0.23</td>
<td>1863560</td>
<td>2256</td>
<td>0.66</td>
<td>1863560</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>20</td>
<td>49279</td>
<td>2839</td>
<td>6.30</td>
<td>317166</td>
<td>28961</td>
<td>9.13</td>
<td>317166</td>
</tr>
</tbody>
</table>

Results for 18 base/expanded query pairs are presented in Table 1. Expansion queries contain from 2 to 34 expansion terms. Note and patient count increases attributable to the expansion queries range from 0.11% to nearly 12.5% and 0.06% to over 6%, respectively.

Discussion

Our results support the assertion that distributional semantic models are a good source of corpus-specific query term synonyms. In this case, the “synonyms” are misspelled orthographic variants. Manual review of the search results indicates false positives are rare. In the case of “acities”, we observed a non-trivial, yet minor fraction of false positives due to the concatenation of ‘a’ and ‘site’ having the same stem as the misspelling ‘asites’. Improvements in recall clearly span a significant range that is not well correlated to the number of expansion terms. In the case of “anesthesia”, expansion term count (27) is at the high end of the dataset, and patient recall improvement is relatively low (0.23%). Yet, given the large number of affected patients, the absolute number of additional patients recalled is quite large. Conversely, ‘serotonin’ has a small number of expansion terms (7) but additional patient recall is rather impressive at 6.2%. Looking at the skewed frequency values in Figure 1, it follows that very common misspellings likely contribute disproportionally to recall. Indeed, ‘serotonin’ alone contributes over 85% of the additional patient recall. Incorporation of DSM-based synonym maps can augment the recall of systems like NLP-PIER through their use in spell correcting user input and by providing users a corpus-specific DSM-based query expansion option.

The last of the ‘serotonin’ expansion terms in Figure 1 and the inclusion of ‘color’ in Table 1 point to the contribution of correctly spelled non-American English substitutions, ‘serotonine’ and ‘colour’, respectively. Magnitudes of such contributions are expected to vary widely; ‘serotonine’ being a minor factor, and ‘colour’ having a major effect on additional recall. Applications of the DSM technique are likely to benefit full text searches where providers contributing to a clinical document corpus hail from a diverse set of countries speaking different languages, e.g., European health care organizations.

Acknowledgements

This research was supported by National Center for Advancing Translational Science (#U01TR002062) (Liu/Pakhomov/Jiang), National Institutes of Health (#8UL1TR000114) (Blazer).

References

Discrepancies in the prevalence of medical conditions based on VA vs. CMS electronic health data

Craig S. Meyer, PhD, MS, MPH, Ning Zhang, MS, Mary Whooley, MD
Measurement Science QUERI, San Francisco VA Health Care System, San Francisco, CA

Objectives: Electronic health records are used to define health conditions used for risk adjustment in health services research, but the accuracy of electronic health records is often unknown. We sought to determine the extent to which predictive model discrimination would improve when data from the Centers for Medicare and Medicaid Services (CMS)\(^1\) was combined with data from the Veterans’ Affairs (VA) Corporate Data Warehouse (CDW).\(^2\) We calculated model performance for 12-month mortality after hospitalization for heart failure using inpatient data files from the VA alone vs. combined data from VA and CMS.

Methods: We used national inpatient files from the VA to identify 44,753 Veterans over age 65 who were discharged from VA hospitals with a primary diagnosis of HF between 2007 and 2011. Among these patients, 11,134 (25%) also had inpatient data available from CMS National Claims History files. For veterans with multiple admissions over the study period, we randomly selected one index admission over the five years. Using scrambled social security numbers, we used a crosswalk to match veterans from the CDW to their Beneficiary ID in the CMS inpatient data files. We compared the prevalence of 14 common health conditions (including hypertension, diabetes, and ischemic heart disease) in the 12 months prior to the admission date of their index hospitalization for HF based on commonly used ICD9-CM code sets for risk-adjustment\(^3\) in VA-only to combined VA-CMS data. Chi-squared test statistics compared the distributions of each health condition separately between the VA-only and the combined VA-CMS data with a significance threshold of \(p<0.05\). Two logistic regression models were used to model 12-month mortality and calculate the C-statistic using each set of comorbidities.

Results: Patients were predominantly male (98%) and White (76%) with a mean age of 77 years. As compared with VA data, the prevalence of health conditions was on average 6.5% (range: 3.9% to 10.5%) higher based on combined data from both VA and CMS. (See Table 1). The C-statistic for the model using VA-only data was 0.598 while the model using combined VA-CMS data had a C-statistic of 0.602 [Difference = 0.004, 95% CI = (-0.001, 0.107)].

Discussion: Among Veterans hospitalized for HF who had inpatient health records in both VA and CMS, the prevalence of medical conditions was higher in combined VA-CMS data files

---


compared to VA-only. The model using combined VA-CMS data had higher predictive performance compared to the model using VA-only comorbidities.

**Impact Statements:** When available, researchers may need to consider multiple sources of electronic health data for accurate risk adjustment variables in health services research.

**Table 1. Prevalence of health conditions in VA vs. CMS data sources for veterans hospitalized with heart failure at the VA, 2007-2011**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Combined Dataset (N=11,134)</th>
<th>CDW Data file (N=11,134)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertension</td>
<td>9,792 (87.9%)</td>
<td>9,093 (81.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Ischemic Heart Disease</td>
<td>8,080 (72.6%)</td>
<td>7,183 (64.5%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Peripheral Vascular Disease</td>
<td>2,281 (20.5%)</td>
<td>1,706 (15.3%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Valvular Heart Disease</td>
<td>3,360 (30.2%)</td>
<td>2,684 (24.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Atrial Fibrillation</td>
<td>5,914 (53.1%)</td>
<td>5,295 (47.6%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Dyslipidemia</td>
<td>6,666 (59.9%)</td>
<td>5,786 (52.0%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Cerebrovascular Disease</td>
<td>1,624 (14.6%)</td>
<td>1,127 (10.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>6,074 (54.6%)</td>
<td>5,483 (49.2%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Depression</td>
<td>1,810 (16.3%)</td>
<td>1,379 (12.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic Obstructive Lung Disease</td>
<td>5,584 (50.2%)</td>
<td>4,419 (39.7%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Chronic Kidney Disease</td>
<td>6,001 (53.9%)</td>
<td>5,161 (46.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Anemia</td>
<td>5,147 (46.2%)</td>
<td>4,358 (39.1%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Tobacco Use</td>
<td>2,107 (18.9%)</td>
<td>1,271 (11.4%)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Implant Cardiac Defibrillator</td>
<td>1,169 (10.5%)</td>
<td>655 (5.9%)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*aAbbreviations: VA = Veterans’ Affairs, CDW = Corporate Data Warehouse, CMS = Centers for Medicare and Medicaid.

*bChi-squared test statistics compared the distributions of each health condition defined from CDW-only to combined dataset.*
Crowdsourcing for Research EEG Annotation and Accuracy Estimation

Andrew Nguyen\(^1\), William Bosl\(^1,2\), Susan Herman\(^3,4\), Tobias Loddenkemper\(^2,4\)

\(^1\)University of San Francisco, San Francisco, California; \(^2\)Boston Children’s Hospital, Boston Massachusetts; \(^3\)Beth Israel Deaconess Medical Center, Boston, Massachusetts; \(^4\)Harvard Medical School, Boston, Massachusetts

**Introduction:** Critical care units increasingly use continuous EEG (cEEG) monitoring to detect seizures and other neurological emergencies in patients who may not exhibit any clinical signs or symptoms due to coma, anesthesia, or trauma. While the value of cEEG has been proven to improve patient outcomes\(^1\), reviewing the EEG record in a timely manner can be quite challenging. Typically, the neurophysiologist reviews the record every 12 hours. EEG Technologists and critical care nurses may be able to augment these readings, but they still cannot provide the kind of continuous monitoring that patients need. Automated algorithms for screening EEG records for epileptiform spikes and other pathological patterns have been proposed as a clinical decision support tool to aid neurophysiologists in monitoring EEG in the ICU. Research on these algorithms is hindered by the lack of sufficient human-annotated EEGs to develop and test learning algorithms, and the lack of standards for consistently scoring the algorithms.

The goal of this project is to create the computational infrastructure to create a research database of annotated EEG records, together with quantitative estimates of the accuracy of all annotations and of each annotator. The proposed crowdsourcing paradigm is widely used outside of medicine in an effort to decrease the cost of labeling data and to increase the speed of labeling, while assigning quantitative estimates of reliability\(^2,3\). The accurate labeling of training data is the cornerstone of any automatic algorithm development. New data may be added continuously, and the system will enable distributed crowds of EEG professionals to engage in training exercises, while simultaneously contributing to annotations. While this will help create a research database of EEG annotations for development of automated algorithms, this project will enable probabilistic quantification of the ‘gold-standard’, along with methodology to quantify the accuracy and expertise of individual labelers\(^4\).

**Methods:** The software infrastructure for this project consists of an interactive web interface to display multichannel EEG signals in a manner similar to commonly used clinical EEG displays. Users may select time windows and choose from annotation label options. The EEG segment to be displayed is taken from a secure database of EEG files, some of which may contain previously attached expert annotations. All annotations from every user are saved in the database, with annotations time-synchronized with the EEG tracing. Metadata about the annotator is also saved, including education, certifications or residency training, skill level, and years of experience. Embedded EEG files with ‘gold-standard’ annotations will be included and randomly presented to the user in order to quantify their accuracy. This information can be recorded with each annotation session recorded for the user.

The goal of this system is to collect performance data of annotators at a very granular level, leading to the quantification of an individual’s accuracy and expertise in EEG monitoring. For each annotation task, we can calculate the overall accuracy of the annotators against the gold-standard. We can also compare the accuracy of each individual annotator against the accuracy of the gold-standard across all tasks. By combining these comparisons with the collected metadata, we will be able to stratify annotators by both quantitative and qualitative measures of accuracy and expertise. The prototype system was built on the Clojure and Clojurescript programming languages. The backend sits atop a column-store database (for signals) and a graph database (for metadata and annotations. All data were de-identified and contain no PHI. The frontend uses a combination of common libraries (React, jQuery, Bootstrap) and the Highstock Javascript chart library (https://www.highcharts.com/). A screenshot of the prototype web interface is shown in figure 1.

An estimate of the true annotation can be obtained by a

![Figure 1. Prototype web interface for annotation.](https://www.highcharts.com/)
consensus vote using algorithms that implement a version of Condorcet’s Voting Theorem. Even in the absence of a ‘gold standard’ annotation, an estimate of the true value can be determined from all annotations if some estimate of the annotator’s accuracy at the task is known. This can be determined by including some previously labeled items that are considered to be ‘gold standard’ in the annotator’s training set. From these, the annotator’s accuracy is estimated for this session and stored. The increase in reliability of additional voters in a crowdsourced labeling task is shown in figure 2. The curves shown in this figure were generated from the Condorcet probability equations:

\[
\begin{align*}
    \text{Prob}(K = k) &= \\
    &= \prod_{i=1}^{n} (1 - p_i), \quad k = 0 \\
    &\leq \sum_{i=1}^{k} (-1)^{k} \text{Prob}(K = k - 1) T(i), \quad k > 0
\end{align*}
\]

where

\[
T(i) = \sum_{j=1}^{N} \left( \frac{p_j}{1-p_j} \right)^i
\]

**Results:** The prototype system allows for any EEG professional to login to the system and begin annotating data. Users are presented with a screen displaying a single page of the multichannel tracing of a random EEG record. The user can zoom in on a page (standard time segment) or scroll through pages to explore the signal. When ready, the user chooses from a list of American Clinical Neurophysiology Society (ACNS) standardized terminologies to add the appropriate annotation. Once set, the user receives immediate feedback showing other annotations of the same record by other annotators, including a consensus vote and the estimated reliability of the consensus vote. If any of the annotations is considered a ‘gold standard’, the accuracy of that may be set to 1.0 and it will fully determine the annotation.

On the backend, a consensus vote is automatically computed when every new annotation is completed. We perform a simple majority vote of all annotations to compare against the ‘gold-standard’ since our initial annotation tasks consisted of the presence or absence of a particular feature. Because the system tracks the time of annotation, users can see their performance against other annotators and against the ‘gold-standard’ as it changes over time. Thus, when the accuracy of annotators is known, an overall consensus can be determined and the accuracy or reliability of that estimate can be computed by the methods that result in the curves in figure 2.

**Discussion:** The prototype system demonstrates that an accessible web portal can be implemented that will enable annotations to be determined with estimates of overall reliability using groups of expert or non-expert annotators. The immediate feedback given to the user enables the system to be used for training, while at the same time producing sets of annotated data that are then available to the community for research.

**Conclusion:** A web-based tool connected to a research database of EEG records has been implemented to enable crowd-based annotation of EEG data with ACNS standardized terminologies. The combination of training, estimating individual accuracy, and production of annotated research data will provide the neurophysiology and neurodiagnostics community with a tool that serves all members on a variety of levels.

**References**

4. Yan, T., Kumar, V. & Ganesan, D. in *MobiSys’10.* (Association for Computing Machinery).
Research Permissions via a Patient Portal: Demographic Profile
Jihad S. Obeid, MD, Azza Shoai, PhD, James C. Oates MD, Melissa A. Habrat, BS, Leslie A. Lenert, MD, MS
Medical University of South Carolina, Charleston, SC

Abstract
We examined the demographic profile of patients in a research permissions registry, containing patient portal users with explicit research preferences on future contact and biobanking. Results show significantly lower odds of volunteerism in minority populations. Although patient portals can be more efficient for recruitment, researchers have to be cognizant of, and proactively address, potential selection bias when using such tools.

Introduction
As healthcare institutions move toward patient-centered care, Personal Health Records (PHRs) play an increasing role with healthcare delivery. As a result, PHRs they are being explored by many academic health centers as means for engaging patients in research both for patient reported outcomes and recruitment into clinical trials. In particular, patient portals, which are PHRs tethered to the patient’s Electronic Health Record (EHR) can be extremely valuable for targeting patient populations with given clinical phenotypes matching research eligibility criteria. However, PHRs may come with inherent non-response selection bias. At the Medical University of South Carolina (MUSC), we have constructed a research registry that is populated using a research permissions questionnaire via the patient portal (Epic MyChart). The questionnaire includes preferences for two permissions: future direct contact for research opportunities and willingness to allow anonymized use of discarded specimens for research (biobank). In this report we present a detailed assessment of the demographic breakdown of our registry.

Methods
We obtained IRB approval to study the demographic breakdown of the patient population, 18 years or older, in our patient portal and research permissions registry during the period between 12/2014 and 3/2016. Basic demographic data on all patients to whom the questionnaire was sent during that period was exported from Epic into SAS software. Patient demographics included: gender, age in years, race, and marital status. Other variables included: ICD 9/10 codes, which were used to calculate the Charlson score; and the status and responses to the research preferences questionnaires. The data was analyzed using SAS software.

Results
Table 1. Breakdown of the 79,837 across different demographic factors and questionnaire responses.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Future Contact for Research</th>
<th>Biobank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>NNR*</td>
</tr>
<tr>
<td>Male (N=27,705)</td>
<td>25.4%</td>
<td>74.6%</td>
</tr>
<tr>
<td>Female (N=52,129)</td>
<td>22.8%</td>
<td>77.2% (Reference)</td>
</tr>
<tr>
<td>Unknown (N=3)</td>
<td>-</td>
<td>100.0%</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-35 (N=21,665)</td>
<td>14.2%</td>
<td>85.8% (Reference)</td>
</tr>
<tr>
<td>35-50 (N=19,095)</td>
<td>21.4%</td>
<td>78.6%</td>
</tr>
<tr>
<td>50-65 (N=22,527)</td>
<td>29.3%</td>
<td>70.7%</td>
</tr>
<tr>
<td>&gt;65 (N=16,550)</td>
<td>31.0%</td>
<td>69.0%</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
</tr>
<tr>
<td>European Americans (N=61,448)</td>
<td>26.8%</td>
<td>73.2% (Reference)</td>
</tr>
<tr>
<td>African Americans (N=14,336)</td>
<td>12.0%</td>
<td>88.0%</td>
</tr>
<tr>
<td>American Indian (N=121)</td>
<td>15.7%</td>
<td>84.3%</td>
</tr>
<tr>
<td>Asian (N=853)</td>
<td>13.6%</td>
<td>86.4%</td>
</tr>
<tr>
<td>Others (N=1,877)</td>
<td>16.14%</td>
<td>83.9%</td>
</tr>
<tr>
<td>Unknown (N=1,202)</td>
<td>20.4%</td>
<td>79.6%</td>
</tr>
</tbody>
</table>

*p-value is less than 0.05. *No, Not Ready or No Response. OR= Odds ratio; CL= Confidence limits.
The total number of patients contacted via the patient portal was 79,837. Of the 25,768 who responded, 18,892 were “Yes” to future contact and 19,713 “Yes” to the biobank. The rest were “no” or “not ready”. Table 1 summarizes the results. Breakdown across the Charlson scale categories (not shown) did not show any consistent trend. Marital status, also not shown, had a small but significant decrease in “Yes” responses (OR=0.88).

**Discussion**
The results show significantly lower odds of opting into both research permissions in minority populations. This is particularly significant in African Americans for future contact for research (OR=0.42) and even more so for opting in for biobanking (OR=0.36), with p<0.05 for both. There were no significant differences across genders; however, there was a significant trend in responses over increasing age categories, with older patients tending to respond more positively towards research opt-in. It is not clear from these results whether under-representation in minority populations in our research permissions registry is due to less frequent adoption of PHRs, distrust in research or a combination of both, the latter scenario being more likely.

Although patient portals can be more efficient for recruiting participants, there needs to be further investigation of the causes of disparities. Further engagement and education in under-represented populations may help allay some of these disparities. In the meantime, researchers have to be cognizant of, and proactively address, potential selection bias when using such recruitment registries especially those involving opt-in approaches through EHR patient portals.

**Acknowledgements**
This work was supported in part by the NIH National Center for Advancing Translational Sciences (NCATS) through Grant Number UL1 TR001450.

**References**
Research Informatics Infrastructure: Landscape and Sustainability

Jihad S. Obeid, MD1, Firas H. Wehbe, MD, PhD2, Paul A. Harris, PhD3, William K. Barnett, Ph.D4, Nicholas R. Anderson, MS, PhD5, Peter J. Embi, MD, MS1, William R. Hogan, MD, MS6, Douglas Bell, MD, PhD7, Leslie McIntosh, PhD8, Umberto Tachinardi, MD, MS9, James J. Cimino, MD10, Peter Tarczy-Hornoch, MD11

1Medical University of South Carolina, Charleston, SC; 2Northwestern University, Chicago, IL; 3Vanderbilt University, Nashville, TN; 4Regenstrief Institute, Indianapolis, IN; 5UC Davis, Davis, CA; 6University of Florida, Gainesville, FL; 7UCLA, Los Angeles, CA; 8Research Data Alliance-US, Troy, NY; 9University of Wisconsin, Madison, WI; 10University of Alabama at Birmingham, Birmingham, AL; 11University of Washington, Seattle, WA

Abstract

A CTSA informatics workgroup recently conducted a landscape analysis designed to understand locus of control and methods of sustainability for component groups (e.g. platforms and services) relevant for supporting research at an enterprise level. The results show some trends in control and sustainability of certain components (e.g. EHR data repositories), but in our assessment there is no fixed blueprint for ownership and sustainability for most component groups.

Introduction

Since its inception, the translational science roadmap proposed by the National Institutes of Health (NIH) highlighted the NIH’s investment in an informatics infrastructure. Academic Health Centers (AHCs) have invested significantly in research infrastructure including research informatics since the beginning of the NIH Clinical and Translational Science Awards (CTSA) program. The need for informatics and data science has become more pronounced in recent years due to the emphasis on using electronic systems to improve efficiency for research studies, data-driven medicine, the learning health system, and precision medicine. However, biomedical informatics support and infrastructure for translational research varies significantly among different AHCs, and there is no prescribed template for what this infrastructure encompasses nor how it should be governed and funded. In this presentation, we highlight the results of an analysis by a workgroup in the Informatics Domain Task Force (iDTF) of the CTSA network focused on identifying the components of a sustainable informatics infrastructure across CTSA funded institutions. The intention is to address gaps in our understanding of the required informatics infrastructure landscape, as well as the different models of governance, funding and support.

Methods

A workgroup of Informatics professionals was assembled through the iDTF to identify the components of a comprehensive clinical and translational research informatics infrastructure and examine avenues of sustainability. The group conducted several meetings and iteratively converged on 16 component groups. This was followed by a survey of the iDTF membership (all funded CTSA) to examine the locus of control (Informatics core, Information technology (IT) department, Research office, or other) and means of funding (institutional, fee-for-service, grants, or other) for each of these components. Respondents were allowed to select multiple options for each component.

Results

Sixty-six individual components were identified by the workgroup and aggregated into 16 groups, under 6 major headings (Table 1). Representatives from 42 of 64 CTSA hubs responded to the survey. Several research components were managed primarily by Informatics, e.g. Electronic Health Records (EHR) data repositories/data warehousing (98% of respondents); Electronic Data Capture (EDC) (93% of respondents); Education and training (93%); and Extramural data collaborations (91%). Virtually all component groups (including the ones mentioned above) had overlapping management by a combination of Informatics, Research office, IT or other. Similarly, all had mixed funding through one or more of institutional funding, fee-for-service, grants or other; however, notably, over 90% of respondents indicated

<table>
<thead>
<tr>
<th>Table 1. The 16 groups of components under the major headings.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Category Heading</strong></td>
</tr>
<tr>
<td>Application for clinical &amp; transnational research</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Research collaboration</td>
</tr>
<tr>
<td>Cyberinfrastructure</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Oversight and governance</td>
</tr>
<tr>
<td>Training and support</td>
</tr>
<tr>
<td>Research and Innovation</td>
</tr>
</tbody>
</table>
that regulatory compliance systems, IT infrastructure, security, EHR data warehousing, governance resources, grants and contracts systems, and clinical trials management systems (CTMS) were all funded primarily by institutional funds. The results for locus of control and sustainability are visualized using radar graphs (Figure 1).

**Figure 1.** a) Distribution responses for locus of control across component groups of infrastructure. b) Distribution responses for sustainability models across different the component groups.

**Discussion**

The results show several component groups that were overwhelmingly supported as institutional infrastructure, albeit under the control of different loci of authority. For example, the IRB systems, grants and contracts, CTMS, data warehousing, and cyberinfrastructure were all considered core components for any AHC engaged in translational research. As such, they are supported by institutional funds as part of operational costs. Grants seem to be the primary funding mechanism for innovation and development of novel research informatics resources such as novel EHR data repositories, self-service cohort discovery tools e.g. i2b2, extramural data sharing collaborations, and basic methodological research.

These data clearly show that there is not a fixed blueprint for a clinical and translational research informatics program. Indeed, there is significant variation of structure, governance, and sustainability for the different components. A better understanding of these factors, through further investigation, may mitigate potential risks to the long-term stability of the infrastructure, and elucidate which sustainability models serve best for return on investment.

The landscape and trends highlighted above are worth noting for AHCs and funding agencies planning for current and future informatics infrastructure, which is the cornerstone for a robust clinical and translational research program and data-driven biomedical research.

**Acknowledgements**

This project has been funded by the National Center for Advancing Translational Sciences (NCATS), grant numbers: UL1TR001450, UL1TR001422, UL1TR000445, UL1TR001108, UL1TR001860, UL1TR001427, UL1TR001881, UL1TR000448, UL1TR000427, UL1TR001417, UL1TR000423, and U54TR000123.

**References**

Extraction of Tobacco Exposure with the Smoking History and Pack-Year Extraction System (SHAPES)

Travis J Osterman, DO, MS, Julie Wu, MD, PhD, Dara E Mize, MD, Wei-Qi Wei, MD, PhD, Josh C Denny, MD, MS
Vanderbilt University Medical Center, Nashville, TN

Introduction

Natural Language Processing (NLP) has been widely used to extract smoking status from clinical notes.1–3 Unfortunately, these systems are insufficient to determine eligibility for lung cancer screening. National screening guidelines require individuals to be 55–80 years of age, have smoked at least 30 pack-years of cigarettes, and have not quit smoking for more than 15 years.4 One pack-year is defined as having smoked 1 pack per day for one year (approximately 7300 cigarettes). To address this deficit, we have developed the Smoking History And Pack-year Extraction System (SHAPES), a rules-based NLP system. SHAPES extracts ever versus never smoking status, smoking rate, smoking duration, total quantity of cigarettes smoked, and years quit smoking from clinical notes.

Methods

To establish a training set, 261 patient records were randomly selected from the Vanderbilt University Medical Center (VUMC) Synthetic Derivative (SD)5, containing 9573 clinical notes. Note identifier; patient identifier; individual date of birth; note date of service; note type; and free, unstructured, text from each note were saved and used for iterative rule creation within SHAPES. A review and annotation system was developed to assist expert reviewers in efficiently labelling and extracting information from patient notes.

Rules for SHAPES were iteratively created by cyclical process of 1) applying SHAPES to a portion of the training set, 2) calculating performance statistics, 3) evaluating failures, and 4) modifying or creating new rules. Cycles were repeated as many times as needed, generally to target an F-measure of 0.90. When the system was performing sufficiently well, the portion of training set included was increased.

A total of 352 patient records containing 4040 notes were randomly selected from the VUMC SD for the validation set. The validation set was reviewed and annotated by two physicians. The two reviewer annotations were compared across each of the six fields. Adjudication was performed on any note that did not have total agreement across all six fields. Adjudication was then performed by a third physician reviewer. The final validation set contained the combination of notes in full agreement and the final adjudicator determination.

Results

Of the 4040 notes in the validation set, reviewers agreed on all extracted smoking variables for 3305 notes (Kappa 0.49–0.91 across six variables). Adjudication was required for 735 notes. Performance statistics for each of the six smoking variables are presented in Table 1. Figure 1 showed calibration plots for SHAPES results compared to gold standard.

<table>
<thead>
<tr>
<th>Neversmokers</th>
<th>Ever-smokers</th>
<th>Rate</th>
<th>Duration</th>
<th>Quantity</th>
<th>Years Quit</th>
</tr>
</thead>
<tbody>
<tr>
<td>N (patients)</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
<td>352</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.79</td>
<td>0.98</td>
<td>0.73</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.98</td>
<td>0.84</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td>PPV</td>
<td>0.95</td>
<td>0.70</td>
<td>0.86</td>
<td>0.71</td>
<td>0.69</td>
</tr>
<tr>
<td>NPV</td>
<td>0.88</td>
<td>0.99</td>
<td>0.94</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.86</td>
<td>0.82</td>
<td>0.79</td>
<td>0.62</td>
<td>0.64</td>
</tr>
<tr>
<td>RMS</td>
<td>n/a</td>
<td>n/a</td>
<td>0.40</td>
<td>6.76</td>
<td>9.99</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.91</td>
<td>0.88</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 1. SHAPES performance versus physician reviewer for ever versus never smoking status, smoking rate, smoking duration, total quantity of cigarettes smoked, and years quit smoking variables. PPV: positive predictive value, NPV: negative predictive value; RMS: root mean square error.
Figure 1. Calibration plots of physician-reviewed tobacco quantity (in pack-years) vs SHAPES prediction (panel A) and number of years quit (panel B).

When Nation Lung Cancer Screen Trial criteria were applied to the validation set, 22 patients were eligible for low dose CT screening for lung cancer. SHAPES identified 17 patients with 1 false positive and 6 false negatives. Based on this validation set, for identifying patients qualifying for lung cancer screening, SHAPES’ precision was 0.94 [0.90,0.96], recall was 0.73 [0.70, 0.75], and F-measure 0.82.

When the United States Preventative Services Task Force for abdominal aortic aneurysm (AAA) screening criteria were applied to the validation set, 35 patients were eligible for AAA radiographic screening. SHAPES identified all 47 patients with 12 false positives and 0 false negatives. Based on this validation set, for identifying patients for AAA screening, SHAPES’ precision was 0.74 [0.73, 0.76], recall was 1.00 [0.95, 1.00], and F-measure was 0.85.

Conclusion

The focus of SHAPES is quantitative smoking data extraction including total cigarette exposure and quit date. The system performs well for pragmatic problems such as identifying patients for lung cancer or abdominal aortic aneurysm screening. Rules can be added without changing the SHAPES’ codebase. Further improvements in individual performance measures may be addressed through iterative expert review and additional rule curation.

References

Has Meaningful Use improved collection of smoking status information?

Fernanda Polubriaginof, MD MA¹, Abdul A. Tariq, PhD², Nicholas Tatonetti, PhD¹, David K. Vawdrey, PhD²,¹
¹Department of Biomedical Informatics, Columbia University, New York, NY; ²Value Institute, NewYork-Presbyterian Hospital, New York, NY

Introduction
Smoking remains the number one cause of preventable death in the United States, responsible for more than 480,000 deaths annually.¹ Obtaining a patient's smoking status is a crucial step in beginning smoking cessation interventions and monitoring progress.² Accurately recording smoking status during a clinical encounter seems like a straightforward task; however this important behavioral determinant of health is often overlooked.³ Given the clinical importance of recording smoking status, the “meaningful use” (MU) financial incentive program for electronic health record (EHR) adoption in the U.S. included a requirement for healthcare providers to capture patients' smoking status electronically in a structured fashion.⁴ To assess whether the data quality of smoking status improved after Meaningful Use requirements were implemented, we reviewed smoking status collected over a 10-year period in a large academic medical center.

Methods
We conducted a retrospective study to analyze smoking status data before and after Meaningful Use criteria were implemented at NewYork-Presbyterian Hospital/Columbia University Medical Center. We included patients that had at least one hospital admission between November 2007 and August 2017. At clinical encounters during the study period, smoking status was collected in a variety of electronic notes authored by several types of providers (e.g., physicians, nurses, social workers). The electronic notes had structured fields for recording smoking status. These fields were standardized in response to MU requirements.

We analyzed changes in the documentation pattern of smoking status between 2007 and 2017. All smoking status observations were mapped to one of four clinically meaningful smoking status categories: 1) “Current smoker,” 2) “Former smoker,” 3) “Never smoker,” and 4) “Unknown smoking status.” Once the categories were mapped, we analyzed smoking status collected over time for each patient and analyzed whether subsequent updates to smoking status were plausible or implausible. The accompanying Figure demonstrates all possible changes in smoking status and labels each change as plausible or not plausible. Additionally, we analyzed the number of discrepancies in smoking status recorded during the same hospital admission for each patient. We assessed quality of smoking status based on the percentage of patients with consistent and informative smoking status recorded in the EHR. We reported the number of patients with smoking status documented, the number of times smoking status was recorded per visit and per patient, the percentage of visits and patients with discrepancies, and the number of plausible and implausible changes. To assess the impact of MU in the data quality of smoking status, we compared the descriptive statistics outlined above during the years pre- and post-MU implementation.

Results
We reviewed data from 304,926 patients having 529,236 hospital admissions during the 10-year study period, wherein 858,512 observations of smoking status were recorded. The accompanying Table presents the number of patients and visits with more than a single smoking status collected, as well as the average number of times smoking status was collected, the number of provider types that collected smoking status, and the rate of discrepancies and implausible changes. Over the 10-year study period, smoking status was more frequently documented. However, the rate of discrepancies increased both in the visit and patient level from approximately 5% to 40%. Similarly, the rate of implausible changes increased from nearly 2% to 31%.

Discussion
The MU program specifies eight distinct categories for collecting smoking status: “Current every day smoker,” “Current some day smoker,” “Former smoker,” “Never smoker,” “Smoker, current status unknown,” “Unknown if ever smoked,” “Heavy tobacco smoker,” and “Light tobacco smoker.” While MU helped to standardize data collection of smoking status, it did not necessarily improve data quality. We observed that the number of times smoking status was collected increased over the years both at the patient level and the visit level. Because the EHR did not provide a
To improve the data quality of smoking status in EHRs, we recommend that patients’ smoking status be stored in a centralized fashion using clinically actionable categories. Clinicians could then more easily verify this information in every encounter by asking patients about tobacco use. To maintain and improve data quality, implausible changes and updates resulting in information loss should require explanation by the user. One way to improve the consistency and correctness of patient-reported information, such as smoking status, is to allow patients to directly review and update their information using kiosks, portals, or printed forms. Eliciting this information via a computer may also mitigate the potential biases introduced by clinicians asking questions regarding smoking behavior.

Figure. Changes of smoking status over time. Dashed lines demonstrate implausible discrepancies and continuous lines represent plausible changes in longitudinal data.

Table. Annual number of patients and visits with smoking status recorded, number of times recorded, number of different provider types recording smoking status, and rate of discrepancies and implausible changes.

<table>
<thead>
<tr>
<th>Year</th>
<th>With smoking status</th>
<th>With more than 1 smoking status</th>
<th>Patients Times recorded (avg)</th>
<th>Provider types (avg)</th>
<th>With discrepancies</th>
<th>With implausible changes</th>
<th>With smoking status</th>
<th>With more than 1 smoking status</th>
<th>Times recorded (avg)</th>
<th>Provider types (avg)</th>
<th>With discrepancies</th>
<th>With implausible changes</th>
<th>Visits Times recorded (avg)</th>
<th>Provider types (avg)</th>
<th>With discrepancies</th>
<th>With implausible changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>6337</td>
<td>0.1379</td>
<td>1.1673</td>
<td>1</td>
<td>0.0395</td>
<td>0.0208</td>
<td>7.042</td>
<td>0.0494</td>
<td>1.0504</td>
<td>1</td>
<td>0.05</td>
<td>0.0197</td>
<td>2008</td>
<td>32788</td>
<td>0.2151</td>
<td>1.3758</td>
</tr>
<tr>
<td>2009</td>
<td>34296</td>
<td>0.2219</td>
<td>1.3784</td>
<td>1</td>
<td>0.0364</td>
<td>0.0266</td>
<td>4.002</td>
<td>0.0476</td>
<td>0.0476</td>
<td>1</td>
<td>0.0704</td>
<td>0.0063</td>
<td>2010</td>
<td>38771</td>
<td>0.3447</td>
<td>1.3913</td>
</tr>
<tr>
<td>2011</td>
<td>42020</td>
<td>0.4321</td>
<td>1.3311</td>
<td>1</td>
<td>0.0104</td>
<td>0.0091</td>
<td>5.4502</td>
<td>0.1247</td>
<td>1.342</td>
<td>1</td>
<td>0.1390</td>
<td>0.0836</td>
<td>2012</td>
<td>42658</td>
<td>0.4604</td>
<td>1.8051</td>
</tr>
<tr>
<td>2013</td>
<td>42864</td>
<td>0.4973</td>
<td>1.937</td>
<td>1</td>
<td>0.1445</td>
<td>0.1639</td>
<td>5.5586</td>
<td>0.2323</td>
<td>1.4937</td>
<td>1</td>
<td>0.2201</td>
<td>0.1552</td>
<td>2014</td>
<td>45912</td>
<td>0.6391</td>
<td>2.7596</td>
</tr>
<tr>
<td>2015</td>
<td>45425</td>
<td>0.6415</td>
<td>2.7917</td>
<td>1</td>
<td>1.6049</td>
<td>0.3644</td>
<td>5.8946</td>
<td>0.3411</td>
<td>2.1513</td>
<td>1</td>
<td>0.4779</td>
<td>0.2859</td>
<td>2016</td>
<td>46193</td>
<td>0.6488</td>
<td>2.8465</td>
</tr>
<tr>
<td>2017</td>
<td>31693</td>
<td>0.6601</td>
<td>2.7326</td>
<td>1</td>
<td>1.6407</td>
<td>0.4015</td>
<td>3.9514</td>
<td>0.6594</td>
<td>2.1917</td>
<td>1</td>
<td>0.3159</td>
<td>0.3159</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

The Meaningful Use program increased data collection of smoking status; however, the quality of the information collected did not improve over time. The rate of inconsistencies and implausible changes in smoking status has risen over the years, challenging the appropriate identification of smokers. Centralized documentation with clinically actionable categories and patient-facing tools might improve the quality of smoking status in EHRs.

References

Expanding the Practice of Reproducible Research in Biomedical Informatics with StatTag and Microsoft Word

Luke V. Rasmussen¹, MS; Abigail S. Baldridge¹, MS; Eric Whitley¹; Leah J. Welty, PhD¹
¹Northwestern University Feinberg School of Medicine, Chicago, IL

Abstract
Reproducible research, a cornerstone of good scientific practice, requires that a manuscript’s analysis code and data can be recomputed by the original investigators themselves or by other scientists. Although software tools exist to facilitate reproducibility, technical barriers preclude their use in the wider biomedical community. To address this, we have developed StatTag – a broadly accessible, free plugin for Microsoft Word that directly connects statistical code and output from R, Stata and SAS to a manuscript.

Introduction
Reproducible research is of growing importance in the scientific community as a mechanism to increase rigor and transparency in the research process¹. The practice of reproducible research, is a comprehensive process beginning with data collection and ending with manuscript preparation and publication². In this work, we focus on the final steps of the reproducible research process, where the ideal end product is a manuscript containing results that can be recreated by an independent third party having access to the original analysis code, data files, and study methods. Existing tools that facilitate reproducible manuscript preparation (e.g., R Markdown, knitr, Sweave and SAS’s Output Delivery System [ODS]), have a steep learning curve, may still require manual transcription of results, or may not support desired document formats – notably Microsoft Word, which remains the mainstay, and sometimes singular option, for manuscript preparation in many fields. Although statistical packages may generate output compatible with Word (e.g., R Markdown, SAS ODS), they require that any proposed changes to the Word document be re-entered into the source code and the manuscript regenerated. We address these challenges with StatTag¹ – a plug-in for Microsoft Word, which takes a novel approach to manuscript preparation.

Methods
StatTag is a plug-in for Microsoft Word, with versions available for Windows 7 and higher, and macOS 10.11 and higher. For Windows, StatTag was developed using the C# programming language, and is compatible with Word 2010 or higher, and Stata 14, SAS 9.4 and R 3.3. For macOS, StatTag was developed in Objective-C as a separate application that interoperates with Word 2016, and is compatible with Stata 14 and R 3.3. In both operating systems, communication with statistical programs is achieved using application programming interfaces (APIs). StatTag (stattag.org) is a free and open-source tool, available under the MIT license.

Results
Once installed, StatTag is integrated into the Word toolbar and allows users to associate code files with a document, annotate (or "tag") the portion of the code file that contains relevant output, and insert that output into a Word document. A user may link more than one code file to a document, and unlike existing tools, may mix statistical programs (e.g., use SAS for primary analyses and R for generating figures). Results are inserted into the document using native Word features – namely fields, text boxes and images. Supported result types include single-value calculations, tables, figures, and direct (“verbatim”) output from the statistical program. After insertion, a user may double-click on the output, allowing them to see the exact code used to generate that value. StatTag also provides an interface to edit statistical code directly from Word, with fidelity to the syntax highlighting found in the native code editor (Figure 1). All embedded output can be individually or collectively updated by re-executing the statistical code in R, SAS and/or Stata. Collaborators are able to modify manuscript text using Word's track changes feature, and the marked-up document (even as a copy of the original) remain connected to the statistical code. Currently, both StatTag and access to the statistical code file(s) are required to view the code associated with a result. However, users may edit the text of a Word document created with StatTag even without having StatTag installed on their machine. As of September 2017, 404 individuals have registered to download the software.

Discussion
StatTag distinguishes itself in the field of reproducible research by supporting multiple statistical programs within a single manuscript document, and reducing barriers to adoption by providing an easy to use graphical user interface
within Microsoft Word. Within the workflow of authoring a manuscript, StatTag supports collaborative authoring such that the statistical code and Word file are connected, but may be worked on independently. With StatTag, modifications to a dataset or analysis no longer require manually re-transcribing results into a manuscript or table, and modifications to manuscript text by collaborators do not require duplicated efforts by re-entering changes in to a source file. Use of native Word objects such as fields, text boxes and images allows formatting of the manuscript using Word’s existing features. Although StatTag was initially developed to support clinical and translational science researchers, it has seen adoption by many domains, including economics and education. Future development of StatTag seeks to further streamline the process of manuscript preparation based on user feedback.

![Figure 1. Code view within StatTag, which allows the creation of new tags used to define statistical output, editing of analysis code, and review of the code used to generate a result.](image)

**Conclusion**
Here we present StatTag as a freely accessible platform to integrate statistical results into Microsoft Word. StatTag provides a user-friendly way to facilitate good scientific practice by supporting reproducible research, and has the potential to change how investigators collaborate with statisticians and analysts on a daily basis.

**Acknowledgements**
StatTag was developed with funding from Northwestern University Clinical and Translational Sciences Institute (UL1TR001422).

**References**
A Cross-Institutional Data Discovery Collaboration: Indexing Institutional Research Data

Kevin B. Read, MLIS, MAS¹, Nicole Contaxis, MLIS¹, Ian Lamb¹, Catherine Larson¹, MLIS, Alisa Surkis, PhD, MLS¹
¹Health Sciences Library – NYU Langone Health, New York, NY;

Introduction

While there are many biomedical data repositories that are designed to help researchers locate specific kinds of data (e.g., GenBank, dbGaP), many research datasets – even data resulting from NIH and NSF funded research – are not shared in these repositories¹. The NIH has devoted considerable resources towards the development of systems to aggregate biomedical research data across multiple repositories to improve discoverability of that data²³, but these systems do not address the discoverability of datasets that have not already been shared through these repositories. The Data Catalog developed at NYU Langone Health is an open source tool for indexing datasets generated by researchers within an institution⁴⁵. This low-barrier approach to data sharing requires only that researchers provide a description of their dataset. The data is made discoverable, while the researchers retain control of the data itself, thus providing an avenue for researchers who are not ready to deposit their datasets in a repository to begin the process of data sharing. Subsequently five other health sciences libraries situated in academic medical centers implemented their own institutional instances of the Data Catalog. The six libraries (from this point on referred to as the Data Catalog Collaboration) meet regularly to share challenges in indexing data from different research domains and meeting researchers’ needs in varied institutional environments.

Methods

NYU Langone Health developed a data catalog to meet institutional needs around data discovery and collaboration. Metadata was developed with institutional needs in mind, while taking advantage of the project lead’s participation in the NIH bioCADDIE initiative to ensure that it was also in alignment with national efforts around data discovery. Collaborations were established at the earliest stages of catalog development with a number of NYU stakeholders with an interest in data discovery and data sharing initiatives.

Data Catalog metadata was developed to maximize the discoverability of datasets across a range of research disciplines while being sufficiently lightweight to not require unrealistic levels of input by researchers or overly time-consuming curation by indexers. The metadata focuses on describing: the types of data being indexed, characteristics of the data (e.g., format, size, equipment used to collect data), funding awards associated with the data, access instructions and restrictions for the data, and contextual information about the dataset (e.g., associated publications)⁶.

To facilitate broader implementation of the Data Catalog across institutions, the code was made open source. The goal of this effort was twofold: 1) increase the discoverability of a larger number of otherwise difficult or impossible to find datasets, and 2) utilize cross-institutional input on Data Catalog metadata, functionality and usability to inform catalog refinements such that the catalog would meet the needs of a broader range of institutions. The code was made available on GitHub and modified to allow for easy implementation and other institutional branding. The availability of the data catalog code was then marketed at professional conferences and through the National Library of Medicine (NLM).

Finally, to align Data Catalog metadata with incoming national data discovery initiatives, the data catalog metadata schema was mapped to the DATS³ schema.

Results

The Data Catalog has been used to index many different types of data. Specifically, it indexes data pull requests from the electronic health record (EHR) for research purposes, and clinical trial, epidemiological, longitudinal, and geospatial population health datasets. Since May 2015 the Data Catalog has 47,350 page views, with 10,148 unique visitors.
A number of stakeholders at NYU have contributed to the development and continued expansion of the Data Catalog, including our Clinical and Translational Science Institute, Department of Population Health and the institution’s clinical data management core. Researcher challenges identified through the Data Catalog’s outreach efforts have also led to the development of a workgroup with institutional data policy stakeholders to address concerns researchers have expressed regarding their lack of knowledge of institutional data policies and data use agreements.

Since outreach began to identify external collaborators looking to implement the Data Catalog, the University of Pittsburgh, the University of Maryland, Baltimore, University of North Carolina – Chapel Hill, University of Virginia, and Duke University have installed the Data Catalog at their respective institutions with funding from the NLM. This collaboration has led to ongoing discussions concerning the refinement of metadata to accommodate new research disciplines at other institutions (e.g., basic science datasets), and improvement of the code as more developers contribute to it.

After mapping Data Catalog metadata to DATS, the project lead has begun preliminary discussions with the NLM metadata team about their data discovery initiatives and the metadata mapping has been shared with them.

Discussion

The Data Catalog is an effective means of indexing institutional research datasets that are otherwise not easily discoverable, while providing a low-barrier way to introduce researchers to data sharing. At NYU Langone Health, the Data Catalog has afforded the opportunity to index datasets that are not discoverable elsewhere, establish institutional partnerships to improve data policy workflows, and eliminate redundancies for data requests. As part of a larger collaboration, the Data Catalog is helping to improve data discovery metadata and usability, and streamline the Data Catalog code for future institutional implementations. As more institutions adopt the Data Catalog locally, it can provide an avenue to uncover institutional research data and inform new developments in data discovery across biomedical research. A future potential benefit of the Data Catalog is that, with metadata aligned with national data discovery efforts, it is poised to allow the sharing of the information about these datasets at a national level, thus allowing the community to decide which datasets are of greatest importance based on access requests. This information on dataset usage can serve to inform NIH priorities for future resource allocation.

References

Design and Development of a Machine Learning Pipeline and Protected Computing Environment

Shelley Rusincovitch, MMCi¹, Charles Kneifel, PhD², Michael Gao, BS², Billy Willis, PhD¹, James Fayson, MIS¹, Stephen Blackwelder, PhD¹, Lawrence Carin, PhD², and Ricardo Henao, PhD²

¹Duke Health, Durham, NC; ²Duke University, Durham, NC

Abstract

Machine learning can be challenging for conventional information technology resourcing in healthcare, especially the requirements of more intense computing, an unfamiliar technology toolkit, rapid interactions with data stores, and the need for associated processes of data governance. In this abstract we describe the process flow for machine learning-based health data science projects, the extension of Duke Health’s protected workspace solution learning, and the collaboration of resources and expertise between our university and health system.

Introduction

Machine learning is a form of modern data analysis based around understanding and prediction. Prediction is a powerful tool in clinical settings, but more conventional information technology (IT) resourcing in healthcare can be challenged by machine learning, especially the requirements of more intense computing, an unfamiliar technology toolkit, rapid interactions with data stores, and the need for associated processes of data governance. Machine learning is an important component of health data science, which also includes computation, data collection, data curation, and software engineering.

In this abstract we describe the process flow for health data science projects based upon machine learning, the extension of Duke Health’s protected workspace solution to the use case of machine learning, and the collaboration of resources and expertise between our university and health system.

Methods

In order to foster innovation in health data science, Duke has deployed six demonstration projects since 2016. These projects include a large breadth of applications including modeling of hospitalizations from a cohort of participants in the Medicare Shared Savings Program (MSSP), optimization of diabetes medication management, and predictions based upon radiology imaging data. Although these projects are in early stages of results, our experiences have been the basis for designing a well-informed use case for computing resource development. We describe the team-based approach in figure 1 and a simplified process flow for the demonstration projects in figure 2.

Duke has previously implemented a protected workspace solution called PACE (Protected Analytics Computing Environment) through the support of the Duke Clinical & Translational Science (CTSA) award and institutional commitments¹. PACE was originally developed to support a broad community of researchers and educators within Duke working with identifiable

Figure 1. The development phase of our demonstration projects is driven by a shared purpose defined by a team-based approach.

Figure 2. A simplified process flow for machine learning-based projects. The development phase (highlighted in color) has been our initial target for protected computing capacity.
health information, and was not specifically targeted for machine learning activity.

Using the health data science demonstrations, we designed and deployed a new use case for PACE infrastructure that included:

- Developing capacity for specialized use cases in high-throughput computing
- Fostering shared resources and leveraging capacity from both Duke University and Duke Health
- Supporting solutions with the right level of protection for considerations including Protected Health Information (PHI) and Intellectual Property (IP) to accelerate progress and efficiency
- Piloting solutions and infrastructure with potential to benefit a larger community of users

**Results**

As illustrated in figure 3, the existing basis of the PACE platform has been extended to incorporate new features targeted to the machine learning use case. The most extensive new capacity has been with the compute nodes, especially graphical processing unit (GPU) resources leveraged from our university-facing research computing infrastructure but deployed within our health system firewall.

The protected network which is used by researchers for their computational work has secure borders and strict limitations are placed on data moving both inbound and outbound from the network. Access to external repositories such as GitHub or similar services are challenging because they could allow the accidental release of protected health information. However, researchers need to be able to use source code and tools from a wide variety of locations.

Our model is to provide a location for the researchers to build their code with broad access to public repositories and then move the resulting objects into the protected network. Our target container for these types of project needs to be simple to manage and include all libraries necessary to run as part of the build process. We have found Singularity, developed at Berkeley Lab, to be an ideal fit. The containers run as the user and so no special security requirements are needed. The containers can be built using any version of Linux that is compatible at the kernel level with the host that will run the container. This means that researchers who may be more familiar with packages built using Ubuntu can run on virtual machines (VMs) that use Red Hat Enterprise Linux, which is more common in enterprise environments. The resulting containers can be signed and moved into the protected network for direct use by the researchers with no additional work.

**Discussion**

To date, four of the six demonstration projects have been deployed in PACE, with the most mature vanguard model achieving the progression from development to testing and clinical acceptance.

Although the initial use case has been focused around the development phase of activity, we intend to design subsequent phases of activity to include testing and deployment phases.

**References**


Figure 3. Features of the PACE Health Data Science use case. Components in solid lines have been deployed, and components indicated in dotted lines are currently under development.
Effects of Site and Method of Administration on Patients’ Preferences for Participation in Research

Azza Shaobi, PhD, Melissa A. Habrat, BS, Degrees¹, Jihad S. Obeid MD, James C. Oates MD, Leslie A. Lenert MD MS
Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC USA

Abstract

This study examines how method/site of administration impacts patients’ permissions for use of surplus tissues in research and for direct contact for research opportunities. We compared the response rate of patients to a self-administered questionnaire adjusted for age, race, ethnicity, and comorbidities through an EHR-patient portal and in person in clinics. Patients were somewhat more willing to contribute surplus specimens for research when asked in a clinical setting (odds ratio 1.21 (95% CI 1.11-1.312).

Introduction

The systematic collection of research preferences empowers patients to make decisions regarding future contact for research and control over utilization of discarded specimens¹. This analysis aims to investigate the difference in positive response rate for research participation using two different methods of elicitation: the first, using a patient questionnaire through an online electronic health record (EHR) patient portal, and the second, using the same questionnaire presented within a clinic’s workflow during an outpatient clinic visit.

Methods

This project was designed to demonstrate the feasibility of implementing a population based EHR-based active opt-in research preferences program at the Medical University of South Carolina¹. The first phase of implementation applied a patient questionnaire delivered through the EHR patient portal. The questionnaire assessed the patient’s preferences for direct future research study contacts (CONTACT) and their willingness to allow use of anonymized excess tissue and fluid specimens (BIOBANK) for research. Patients who are portal users received a series of three emails about one month after portal enrollment, then annually, requesting their response to the questionnaire. Then they log into their portal account and complete the online questionnaire at their convenience. In the second phase, the same questionnaire was presented within a clinic workflow, between rooming of the patient and the provider visit using the exam room computer. In this phase, we rolled out on a clinic by clinic basis, a protocol and software tools in Epic designed to support direct patient responses to questionnaires in the period between rooming of the patient and the provider visit. A drop-down menu item opens a new session in MyChart with the patient’s questionnaire queue displayed. When upon finishing the questionnaires, the session’s window closes and the screen is locked and ready for the provider. Providers may also interrupt and take over the computer before completion of questionnaires to preserve work flows. In this test, the research preferences questionnaire (2 items) was the only questionnaire in the patient queue. Permissions from both phases were tabulated in a registry for use by investigators for feasibility assessment of research studies and recruitment. Using this registry, we compared the responses of patients who responded via the patient portal with those who responded in clinic. We used a multivariate logistic regression to compare the odds of positive response among patients who submitted their responses, adjusted for co-variates for positive responses. Models were fitted separately for the biobank and the contact questions.

Results

We analyzed data from responses submitted between April 1st 2016 and April 30th 2017. During this period 16,494 unique patients’ responses were identified. Of those, of 12,193 (73.9%) were submitted through the EHR-portal and 4301 (26.1%) were submitted through the in-clinic workflow method. The in-clinic submission rate followed a linear increasing trend with time while the EHR-portal showed a seasonal trend related to annual renewal requests. After training of staff, in clinic response rates were high and the workflow was well accepted. We obtained similar overall positive response rate from patients who used the clinic workflow method and from those who used the patient portal; 3008 (69.9%) patients responded positively on the biobank question, and 2802 (65.2%) on the CONTACT question through the in-clinic methods, compared to 8544 (70.1%) for biobank, and 8180 (67.1%) for
contact through the patient portal. (see table 1). Demographic and health factors such as age, race and chronic illness were also associated with the odds of a positive response (see Table 1). Older patients were more likely to respond positively and a linear trend is observed. Females were less likely to respond positively for BIOBANK permissions but not CONTACT permissions. African American patients were less willing to be contacted for research (odds ratio 0.57, 95% CI 0.52-0.23) and largely unwilling (odds ratio 0.35, 95% CI 0.32-0.38) to allow surplus specimens to be used for research. Patients with lower Charleston scale (healthier) were less likely to respond positively to either request when compared to those with the highest scale. After controlling for all covariates, the odds of responding positively using the clinic submission were higher for biobank permissions (1.21 (95% CI; 1.11-1.31)) and similar for 1.00 (95% CI; 0.92-1.08) CONTACT permissions when administered in person.

| Table (1): Adjusted Odds Ratios for permissions for direct contact and surplus tissue use for research |
|-------------------|-----------------|----------------|-----------------|
| Factor             | Contact OR      | 95% CI         | Biobank OR      | 95% CI          |
| Age (reference age >65) | 0.76            | (0.68-0.83) *  | 0.74            | (0.67-0.82) *   |
| Gender (Reference = Male) |                |                |                |                |
| Female             | 0.93            | (0.87-1.01)     | 0.85            | (0.79-0.92)     |
| Race (Reference = White) |                |                |                |                |
| African Americans  | 0.57            | (0.52-0.63)*    | 0.35            | (0.32-0.38)*    |
| Chronic illness (Reference = Charlson scale 3) |                |                |                |                |
| Charlson scale 0   | 0.70            | (0.51-0.95)*    | 0.50            | (0.35-0.71) *   |
| Charlson scale 1   | 0.95            | (0.69-1.31)     | 0.73            | (0.50-1.05)     |
| Charlson scale 2   | 0.82            | (0.54-1.26)     | 0.58            | (0.36-0.92)     |
| Ethnicity (Reference = Hispanic) |                |                |                |                |
| Not Hispanic       | 0.96            | (0.69-1.33)     | 0.89            | (0.63-1.25)     |
| Method (Reference = MyChart submission) |                |                |                |                |
| Clinic Submission  | 1.00            | (0.92-1.08)     | 1.21            | (1.11-1.31)*    |

*Statistically significant at a confidence level of 0.05

Discussion

Within the limitation of this analysis, our results indicate it is possible to establish a research permissions registry using the patient portal with a patient-centric, opt-in approach that combines in clinic and portal based administration. In clinic administration may make patients somewhat more likely to agree to have surplus tissues used in research. The in-clinic approach provides the same opportunity to patients who do not use the patient portal, and is associated with a statistically higher positive response rate for permission for surplus tissue use. Age, gender, race and health have to be considered in understanding opt in rates for research participation. As previously seen older age, white race and pre-existing illness pre-disposed to positive permissions. Low consent among African Americans is of particular importance given its effect on the generalizability of research results and health disparities. Further research is required to understand the role of race.

Documenting research participation preferences in the EHR offers the advantage of 1) linking these preferences to health record information for recruitment, 2) the ability to integrate preferences into population based patient registries, and 3) offering automate notifications of surplus specimen availability in particular patient phenotypes.

Research preferences are an important aspect of patient preferences that can be potentially integrated with other patient-reported outcomes and engagement data collection efforts.

References


Acknowledgments: This work was supported by the South Carolina Clinical & Translational Research (SCTR) Institute, with an academic home at the Medical University of South Carolina, NIH/NCATS Grant number UL1 TR001450.
Vivli: A Global Secure Data Sharing Platform for Participant-Level Clinical Trial Data
Ida Sim, MD, PhD,1,2 Rebecca Li, PhD,2,3 Simona Carini, MA,1,2 Pablo Gazmuri, BA,4 Chris Mavergames, MLIS,5 Barbara Bierer, MD3
1University of California San Francisco, San Francisco, CA, 2Vivli Center for Global Clinical Research Data, Cambridge, MA, 3Harvard Multi-Regional Clinical Trials Center of Brigham and Women’s Hospital & Harvard, Cambridge, MA, 4Blue Metal, Inc, Watertown, MA, 5Cochrane, Freiburg, Germany

Introduction and Background
Clinical trials are one of the most important sources of evidence for translational research and care. Their data hold great potential to unlock discoveries for the advancement of human health. Transparency of clinical trials has evolved over the years from trial registration to required reporting of summary-level results data. For greatest value, individual participant-level data (IPD) should also be shared to enable new discoveries beyond those possible with only summary-level data, to validate findings by making it possible to re-analyze and review prior results, and to prevent unnecessary repetition of trials — and the risk to additional patients — when IPD already exist.

To date, IPD sharing has been limited by problems with capacity, policy, technology, and resources. Although there are over 60 sites that currently offer IPD for sharing, they have different standards, security, and access policies, and there is no way to search across all the sites. Because these sites serve limited constituencies (e.g., Clinical Study Data Request (CSDR) hosts only studies from some pharmaceutical companies, BioLINCC hosts only studies funded by the National Heart Lung and Blood Institute, WWARN hosts only malaria-related studies), they provide only a small fraction of the total data sharing capacity needed for IPD sharing to become the global norm.1 Moreover, there is no capability for either federated query or for data from disparate sources to be brought together securely for analysis. Resources, both human and financial, are constrained and sustainability models are unclear.

In a series of meetings in 2013-2016, the Brigham and Women’s Hospital and Harvard Multi-Regional Clinical Trials Center (MRCT Center) convened relevant stakeholders to consider the most significant challenges to realizing large-scale IPD sharing. The need for a new global, secure, neutral platform to share anonymized IPD was identified. As a result, Vivli was launched as an independent not-for-profit organization in the fall of 2016 as a permanent vehicle for the development, management, and governance of a global clinical research data-sharing platform. We present the technical design and associated data use policies for the Vivli data sharing platform.

Methods
Under the MRCT Center, we set up IT and Governance Working Groups comprised of members from academia, non-profits, and the pharmaceutical industry. We reviewed the technical and policy approaches of related initiatives in data sharing, data curation, and data request review, from both industry and academia/non-profits (e.g., Cochrane Collaboration).

The platform design process started with enumerating the main target users and use cases. Technical and policy decisions were made in parallel as needed. For example, some trial investigators supported unrestricted downloading of anonymized IPD while others required a data request review process after which approved IPD must be provided in a secure virtual machine. We iterated a technical approach that will support and enable both current policy needs as well as a longer-term vision of a robust data sharing culture. A draft technical blueprint was presented publicly at a meeting in London, UK in March 2016. In response to feedback from the meeting and guided by use cases, we interviewed data submission and data request users and developed a wireframe prototype.

Together with development partner BlueMetal, an Insight company, we then defined and implemented the system architecture and security. Together with metadata curation partner The Cochrane Collaboration, we defined a metadata curation and annotation strategy. Vivli aims to serve the international community including studies from any disease, country, sponsor, funder or investigator.

Results
The Vivli platform was constructed using modern web and cloud technologies on Microsoft’s Azure cloud platform.
The frontend was built with React, Typescript and dotnet core, while the backend uses Azure cloud services including CosmosDB, Azure Search, KeyVault, Encrypted Storage, Active Directory B2C, Service Bus and IaaS VMs. Data security is built into the system from the ground up: separation of concerns and adherence to principle of least privilege provides defense in depth, hardware security modules are used to secure system secrets, all data within the Vivli platform is encrypted in transport and at rest, and all database updates are audited and logged with digital signatures that include user tokens signed by Microsoft’s Azure Active Directory B2C identity management system, ensuring a high level of assurance and auditability of both user identity and integrity of backend data.

We defined core user roles: e.g., Data Provider, Data Curator, Data Requester and Review Panel Reviewer. An individual can play different roles at different times. For example, a university researcher can contribute one of her studies to Vivli and also search the Vivli Platform to identify and request IPD of other studies.

We prioritized scientifically accurate and precise searching for relevant IPD. The Vivli search interface supports both plain keyword as well as PICO (Participant-Intervention-Comparison-Outcome)-based queries. All studies available on Vivli are described by a metadata model that is based on a merging of the Cochrane ontology and the Ontology of Clinical Research. Annotation terms are drawn from the Cochrane Vocabulary. The process for metadata curation is as follows: a Data Provider submits a trial register ID (e.g., ClinicalTrials.gov NCT ID) and, optionally, the final version of the study protocol. Using the NCT ID, Vivli programmatically queries ClinicalTrials.gov through its API and extracts administrative metadata elements (e.g., study title, study start and completion dates, country of study sites). A curator then manually extracts the PICO elements from the registration record or study protocol and annotates them to Cochrane Vocabulary terms. After passing a quality assurance step, the metadata record is then posted on Vivli, making the study available for search and IPD request. To date, 27 studies have been curated. A video demonstrating the current search interface is available at http://vivli.org/.

Vivli offers a common data request form for users to detail their planned research question and analyses. Completed forms are routed to the appropriate review process for the requested study. For approved requests, the anonymized IPD sets are made available on a secure virtual machine (VM) where data from different hosts and data generators -- including industry, academia, and biotech -- can be aggregated and analyzed with R, STATA, SAS, and other analysis tools. In this controlled-access environment, strict data security models and export controls are applied according to the specifications of each data contributor. Once analyses are complete and their results downloaded, the VM is deleted. All study and derivative datasets are issued DOIs. Long-term data archiving on Vivli is an option.

Discussion

Vivli is poised to launch in July 2018 with the most complete IPD data-sharing platform to date. Through a PICO-based search portal and unprecedented attention to accurate metadata, Vivli will offer a user-friendly one-stop search across multiple IPD data hosts, while providing researchers streamlined and harmonized data request and access policies. Implementation in the Azure cloud allows the highest-level security, and supports anticipated future deployment of machine learning approaches to improve metadata curation and search performance. Over time, Vivli will also adopt technologies to address needs such as anonymization and data mapping, and will expand the types of research data hosted (e.g., observational studies, genomic data). Cost and charge models are under development.

IPD sharing makes the best use of society’s investment in clinical research, and honors the contributions of clinical research participants. Vivli is a ground-breaking initiative to enable secure, trusted IPD sharing on a global scale.

Acknowledgements

Major financial support comes from the Laura and John Arnold Foundation. Microsoft Corporation is a supporting partner. We acknowledge the support of the Brigham and Women’s Hospital and Harvard Multi-Regional Clinical Trials Center, and the contributions of the MRCT Center Data Sharing Working Groups.

REFERENCES:

Influence of bias introduced by filtering patients with “complete” electronic health records (EHRs) on predicting diabetic kidney disease

Xing Song, PhD1, Lemuel R. Waitman, PhD1, Yong Hu, PhD2, Alan S.L. Yu, MD3, David Robins, MD4, Mei Liu, PhD3,
1Division of Medical Informatics, University of Kansas Medical Center, Kansas City, KS, USA; 2Jinan University, Big Data Decision Institute, Guangzhou, PRC; 3Division of Nephrology and Hypertension and the Kidney Institute, University of Kansas Medical Center, Kansas City, KS, USA; 4Diabetes Institute, University of Kansas Medical Center, Kansas City, KS, USA

Introduction
Diabetic kidney disease (DKD) is one of the most frequent and dangerous complications of Diabetes Mellitus (DM), affecting about 40% of the DM patients. Even a mild degree of albuminuria and decrease in estimated glomerular filtration rate (eGFR) are associated with significantly increased risks for cardiovascular disease and death and higher healthcare cost1. Thus, it is extremely important to identify patients at higher risk for developing DKD and implement appropriate interventions early. The nationwide adoption of EHR enables researchers to build risk prediction models for DKD based on more diverse patient populations.

However, EHR data is often incomplete due to various reasons: a data point was observed but not recorded; a data point was never clinically necessary to be observed, or the same patient received care at multiple institutions. The mere availability of a record for a given patient does not necessarily mean that it contains sufficient information for developing accurate predictive models. As a result, investigators may decide to only include patients whose data are “complete enough” for their study. For example, a study may require medication or laboratory tests results for its analysis cohort but not other information. When different data completeness filters are used to sample patients with sufficient information, potential biases can be introduced in the resulting cohort and subsequently affect the downstream analysis (e.g. predictive modeling). This study aims to investigate the effect of different data completeness filters on the study cohort and subsequent prediction of DKD risk among patients with type 2 DM.

Methods
A retrospective cohort of 38,572 DM patients (age ≥ 18) was extracted from our institution’s de-identified data repository HERON2, among which 18,017 developed DKD later. DKD patients were identified by albumin-to-creatinine ratio (ACR) ≥ 30 mg/g or estimated Glomerular Filtration Rate (eGFR) ≤ 60 mL/min/1.73m². For each patient, we extracted demographics, family history of kidney disease, smoking history, and four types of clinical variables including common vital signs, a selective list of diagnoses of comorbidities (hypertension, cardiovascular diseases, hyperlipidemia) relevant to renal insufficiency which have been established in literature, specific medications for treating the comorbidities and controlling glucose level, as well as laboratory tests reflecting kidney function (ACR, eGFR excluded), glucose and cholesterol level. We evaluated three groups of heuristic filters echoing the ones used by Weber et al3:

1. Filters based on patient demographics, which included two filters checking whether the patients have a recorded race (Race) or both age and sex (AgeSex);
2. Filters based on availability of clinical facts, which included four pairs of filters (VitalAny, VitalAll; DiagAny, DiagAll; MedAny, MedAll; LabAny, LabAll) that selected patients with “any” or “all” of the clinical data of interest for each type (predictors used for defining the filters are detailed in Figure 1);
3. Filters based on recency of clinical facts, which included four filters (30d, 60d, 180d, 365d) that excluded patients who have not visited our healthcare facility in a certain temporal proximity to the event of interest.

For better result interpretability, we employed LASSO logistic regression as the common learning model across filters. The area under the receiver operating characteristics (AUC) curve is reported for evaluating prediction performance on 10-fold cross validation. Since the DKD prevalence rate is different after applying each filter, to ensure fair comparison across filters, we performed 1:1 patient matching on their demographics.

Results
As detailed in Figure 1, the demographic information is well populated in EHR, which has only filtered out less than 3% patients, mostly due to unreported “race”. Among the data type filters, 78%, 69%, 65%, and 61% patients passed the MedAny, LabAny, VitalAny and DiagAny filters respectively. However, LabAll, MedAll and DiagAll suffered
from a major cut in their cohort size due to rarity of some categories within each type (e.g., only 4% had been prescribed with fenofibrate), leaving a few very sick patients. In contrast, 60% patients made through the VitalAll filter, as a result of the vital signs being commonly checked and well tracked among DM patients. Almost all the filtered population biased mildly toward non-DKD except for MedAll and LabAll.

Figure 1 – Filtered Cohorts Descriptions

All predictors used in the study were selected by the LASSO algorithm at the end regardless of filters, but their contributions differ considerably resulting in AUC deviations. As in Figure 2, learning from cohorts filtered by DiagAny, LabAny, VitalAll, 30d, 60d and 180d provided significantly better AUC than the others, even some of them were based on only less than half of the overall data. For example, the 60d filter gained 0.021 in AUC compared to MedAny (p-value < 0.001). On the other hand, AUC ranged modestly from 0.64 to 0.68 across filters, which implies that “data completeness” may not be a big issue in predicting DKD if we adopt learning models that can be as robust as LASSO logistic regression against missing data. Greater improvement in performance may be gained with future exploration of more comprehensive list of predictors or predictive models.

Figure 2 – Retain Percentage vs. AUC [95% Confidence Interval]

Discussion

For the specific task of predicting DKD risk among DM patients, we observed that different data completeness filters do affect AUC significantly. On the flip side, if optimal level of data completeness cannot be achieved, the current EHR data may still be adequate to generate reasonable predictions if robust learning models are applied.

Acknowledgement

This research was partially supported by the Major Research Plan of the National Natural Science Foundation of China (Grant No. 91746204, No. 2017B030308008, No. 603141789047). The HERON data repository used in the paper is supported by the University of Kansas Medical Center and CTSA grant UL1TR000001 from NCRR/NIH.

References

**EHR-Integrated, Machine Learning-Driven SMART on FHIR Pharmacotherapy Decision Support System for Type-2 Diabetes Mellitus**

Wataru Takeuchi, MS¹, Shinji Tarumi, MS¹, Salvador R. Loya, PhD², David E. Shields², Phillip B. Warner, MS³, Michael Flynn, MD², Kyle M. Turner, PharmD², Farrant H. Sakaguchi, MD, MS², Hideyuki Ban, PhD¹, Kensaku Kawamoto, MD, PhD, MHS²

¹Hitachi Ltd., Kokubunji, Tokyo, Japan; ²Univ. of Utah, Salt Lake City, Utah, USA

**Introduction**

Type-2 diabetes mellitus (T2DM) is a chronic disease resulting in significant economic, psychological and physical burden to patients. T2DM is typically managed through lifestyle changes as well as oral and/or injectable medications. The established treatment target for most patients is controlling hemoglobin A₁c (HbA₁c) levels to less than 7%; however, about half of patients fail to achieve this target. To help address this important public health problem, the authors have developed a machine learning based pharmacotherapy decision support system (PDSS). In order to maximize the potential for scalable deployment, this PDSS uses standards-based clinical decision support (CDS) interoperability frameworks, including (i) the Fast Health Interoperable Resources (FHIR) Standard for Trial Use (STU) Release 2 data exchange interface developed by Health Level Seven International (HL7); (ii) the open-source OpenCDS CDS Web service framework, which complies with the HL7 FHIR Clinical Reasoning decision support service interface; and (iii) the Substitutable Medical Apps Reusable Technologies (SMART) framework for integrating apps into electronic health record (EHR) systems. A central component of the PDSS is a module encapsulated within OpenCDS which predicts the outcome of prescribed medications, as the established clinical guideline in this area provides significant optionality in pharmacotherapy regimens, leading to a suboptimal trial-and-error approach to drug selection. The aim of this study was to develop a PDSS with good predictive power of which pharmacotherapy regimen would have the highest probability to determine achievement of treatment goals (defined as an area under the curve [AUC] >= 0.80) and to evaluate the feasibility of improving T2DM care by using the PDSS.

**Materials and Methods**

The PDSS consists of the predictive model, encapsulation of the predictive model by OpenCDS, and a SMART on FHIR Web-based dashboard which enables physicians to review relevant patient parameters, select treatment goals, select alternate medication treatment strategies, and review the likelihood of achieving the goals for each treatment strategy. This dashboard uses the FHIR STU2 application programming interface (API) to obtain data and SMART to integrate with the EHR user interface (Figure 1).

![Figure 1. Overview of Pharmacotherapy Decision Support System (PDSS)](image)

To develop the prediction module, we analyzed a dataset of primary care clinic patients at the University of Utah with type 2 diabetes mellitus. The numbers of cases contained in this dataset were 16,545 and 6,281 where post-treatment HbA1c measurements were available after 90 and 180 days, respectively. The ratio of cases whose HbA1c were more than 7% at the treatment decision point in this dataset was 58%. The dataset encompassed 2012 through 2016 (5 years). The data was divided into training data (70%) and validation data (30%). We developed a prediction model which predicted the probability of achievement of two treatment targets: controlling HbA1c to less than 7.0%...
at 90 days after prescription and controlling HbA1c to less than 7.0% at 180 days after prescription using a machine learning method (gradient boosting\textsuperscript{8}) on training data. This prediction model is based on an ensemble of many decision trees. This prediction model uses 40 input variables including the patient’s demographics, vitals, and laboratory tests, such as age, body weight, HbA1c level, blood pressure, and triglyceride level. This prediction model also uses current medication classes (e.g. insulin, metformin, sulfonylurea, and various combinations) as input variables. To evaluate the performance of the developed outcome prediction model, we conducted 5-fold cross validation for hyperparameters tuning and one external validation using the previously extracted validation dataset. The accuracy of the prediction and the area under the curve (AUC) of the receiver operator curve (ROC) were evaluated. The outcome improvement ratio, which is a ratio of the number of patients who achieved the treatment target with the medication recommended by this prediction model compared to one calculated from actual records, was also evaluated. The recommended medication consisted of the medication with the highest probability of achieving a treatment target of 7.0% within 90 or 180 days of therapy initiation.

This study was approved by the institutional review boards of both the University of Utah and of the Research & Development group of Hitachi, Ltd.

Results

The accuracy and AUC of the 90-day outcome prediction model in 5-fold cross validation on training data were $0.808 \pm 0.007$ SD and $0.876 \pm 0.006$ SD, respectively. The accuracy and AUC of the 90-day outcome prediction model on the validation data were 0.814 and 0.882, respectively. The accuracy and AUC of the 180-day outcome prediction model in 5-fold cross validation on training data were $0.822 \pm 0.017$ SD and $0.897 \pm 0.010$ SD, respectively. The accuracy and AUC of the 180-day outcome prediction model on validation data were 0.832 and 0.904, respectively. The outcome improvement ratio on validation data were 1.22 and 1.16 for 90 and 180 days after prescription, respectively. The PDSS utilizing this prediction model was successfully integrated with the Epic\textsuperscript{®} EHR using SMART on FHIR and OpenCDS, and iterative enhancements have been made based on clinician feedback. Formal usability testing and a prospective clinical trial are currently in planning.

Discussion

A machine learning based pharmacotherapy decision support system (PDSS) for T2DM was developed and successfully integrated with the EHR through standards-based interoperability frameworks including FHIR and SMART. This integration of the PDSS with the EHR should facilitate its clinical utility and impact. The prediction model had good predictive power and the potential to improve patient outcomes if recommended regimens were followed. A prospective trial is in planning for evaluating the impact of the PDSS on patient outcomes.

Limitation

This study was a retrospective study. Therefore, the actual effectiveness of this PDSS requires prospective evaluation.

References

7. HAPI-FHIR. Available from: http://hapifhir.io/
Identifying Biologically Implausible Anthropometric Values in a Multisite Extraction of EHR Data for Secondary Use

Daryl L. Wieland, MD, MS1, Caroline S. Jiang, MS2, Amanda Cheng, MPH3, Peter R. Holt, MD2, Jan L. Breslow, MD2, Dena Moftah, BA3, Jonathan N. Tobin, PhD2,3
1NYC Health + Hospitals/Jacobi & North Central Bronx, Department of Ob/Gyn;
2Rockefeller University, New York, New York;
3Clinical Directors Network, Inc. (CDN), New York, New York

Introduction

The fundamental purpose of electronic health record (EHR) data collection is to help clinicians manage an individuals’ health at the point of care and longitudinally; however, the digitalization of information within EHRs provides the opportunity to use the data for other purposes including sharing with other providers, advancing the quality of care delivered to populations, and providing research opportunities to evaluate health outcomes. For each of these uses, the quality of data is essential.

Anthropometric data (height, weight, and body mass index (BMI)) are some of the most common data collected during a routine clinical encounter especially in primary care. Despite the underlying importance of the accuracy of this information, errors in both the ascertainment and recording of these data have been identified, which contribute to patient safety concerns for individuals and may be later recognized during research studies by statisticians who use a variety of approaches to clean and validate extracted data. Common errors include data-entry type errors and unit conversion errors between US Standard and metric. Many of these errors result in the recording of biologically implausible values (BIVs), which could be recognized at the time of data entry with support of technology. Understanding the rates of these errors could alert EHR vendors and informaticians to develop technical solutions to prevent these errors from occurring at the point of care and improve data quality for the care of the individual and for population studies.

In this multisite study of Federally Qualified Health Centers (FQHCs) and community hospitals in an urban, low resource setting, we aggregated five years of data from five centers with different EHRs. Using previously described frameworks to assess the data quality, we quantified the rate of data exclusion based on missing values and biologically implausible values to highlight variation in data validity in the course of clinical care.

Methods

We created a community-academic partnership that included New York City FQHCs (n=4) and hospitals (n=4), The Rockefeller University, The Sackler Institute for Nutrition Science and Clinical Directors Network (CDN – www.CDNetwork.org), a primary care practice-based research network (PBRN). Using the Community-Engaged Research Navigation model, we developed a multisite de-identified database extracted from EHRs of female adolescents and young adults aged 12-21 years in Jan 2011 – Dec 2012. The first round of data extracted included age of patient at entry of study, demographic information, visit type, height, weight and BMI. Data on height and weight were received from the sites as input at the point of care in units in either US standard (feet, inches, pounds) or in metric (centimeters, kilograms), and BMIs were calculated both within the EHRs as well as separately in the analysis dataset. All data were converted to US standard for analysis. This study was reviewed and approved by respective IRBs for each site.

Statistical analysis was performed using SAS Studio 3.6. Univariate statistics with histograms and boxplots were examined for height, weight and BMI. The ranges were set to exclude height <50 inches or >79 inches, weight <=52 pounds or >440 pounds and BMI<12 kg/m^2 or >80 kg/m^2. These values were selected based on Centers for Disease Control reference values for girls age 12-20 for 2000 and minimum and maximum values established by Freedman et al1. The data were cleaned to remove implausible values.

Results

Data extracted from the initial five sites which submitted the first round of data were analyzed. This included 85,115 unique patients with a total 410,861 recorded visits between 2011 and 2015. Not all recorded visits had data for height, weight, or BMI. One site (site E) only provided height and weight in the data transmissions citing difficulties with the extraction process of their database. After the removal biologically implausible values, the means for height and weight were minimally affected, the mean for BMI was dramatically reduced, and the standard deviations for height, weight, and BMI were decreased significantly (Table 1). The cleaning revealed rates of implausible values for height, weight, and BMI of 1.39%, 0.20%, and 1.84%, respectively, across all sites. Exclusion rates of values varied by site, from 0.07% to 4.31%, with the greatest numbers of exclusions in the BMI values.
<table>
<thead>
<tr>
<th></th>
<th>Before Clean</th>
<th>After Clean</th>
<th>Before Clean</th>
<th>After Clean</th>
<th>Before Clean</th>
<th>After Clean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (inches)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight (lbs)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMI (kg/m²)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N OBS</td>
<td>355,408</td>
<td>350,469</td>
<td>368,373</td>
<td>367,642</td>
<td>304,093</td>
<td>298,488</td>
</tr>
<tr>
<td>MEAN</td>
<td>63.15</td>
<td>62.96</td>
<td>149.09</td>
<td>148.69</td>
<td>48.14</td>
<td>26.3</td>
</tr>
<tr>
<td>STD DEV</td>
<td>100.93</td>
<td>3.07</td>
<td>221.97</td>
<td>42.99</td>
<td>2,591.9</td>
<td>7.14</td>
</tr>
<tr>
<td>MINIMUM</td>
<td>0</td>
<td>50</td>
<td>-135</td>
<td>52</td>
<td>-25.5</td>
<td>12</td>
</tr>
<tr>
<td>MAXIMUM</td>
<td>42,836</td>
<td>79</td>
<td>132,132</td>
<td>440</td>
<td>907,000</td>
<td>80</td>
</tr>
<tr>
<td>N EXCLUDED by applying cutoffs</td>
<td>4,939</td>
<td>731</td>
<td>5,605</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% EXCLUDED by applying cutoffs</td>
<td>1.39%</td>
<td>0.20%</td>
<td>1.84%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Descriptive Statistics of combined sites before and after applying cutoffs for implausible values

To further evaluate the errors in the BMI, the BMI was calculated by applying the CDC BMI formula for adults on the extracted height and weight values and compared to the BMI values recorded in the EHR. If the BMI is calculated correctly by the EHR then the difference between these values (BMI-diff) should be 0. The difference in the calculated BMI and the BMI recorded in the EHR showed the greatest variation in Sites B and D (Figure 1). These findings suggest the likelihood of a programmed calculation error within the EHR.

Discussion

This analysis of the quality of routine clinical anthropometric data extracted from the EHR by eliminating biologically implausible values based on using the 1st to 99th percentile from CDC and the NHANES evaluation of weight and weight for pediatric populations revealed that the EHRs from five community sites allowed frequent inaccurate data entry of heights and weights. In addition, the error rates in the BMI calculation suggest that the equation for calculating BMI may have been programmed incorrectly as it may be applying the formula to parameters measured in different units.

Data entry errors are known and not uncommon, but configuration of the EHR to include error checking at the point of input could prevent implausible values from being entered by providing alerts to the user. Entered values for height and weight could be compared with age and sex based norms as well as prior values for a given patient, displaying data in both US standard and metric and properly calculating the BMI. Periodic quality reviews with IT analysts could assess the success of such efforts.

This study has several limitations. Although we reviewed with clinicians anecdotally the possible cause of error at point of care, we were unable to review individual records to quantify the types of errors. This study does not describe the details of the EHR workflow that may have contributed to errors.

Despite recognition of problems associated with the quality of EHR data available, errors still occur in the entry of anthropometric data, numerical data with known biologically plausible values. Efforts should be undertaken to improve standardization of data entry and checking at the point of care to improve safety for individual patients and the quality of data for secondary use for population health impact, outcomes studies, and comparative effectiveness research.

References

Evaluating a Practice-based Health IT Intervention Using Audit Log, Qualitative and Clinical Data

J.B. Jones, Ph.D.1, Xiaowei Yan, Ph.D.1, Jake Delatorre-Reimer 1, Hannah Husby, MPH 1, Cory Mosser1, Ridhima Nerlekar M.S.1, Shuting Liang, MPH1, Satish Mudiganti, M.S.1, Zijun Shen, M.S.1, Walter F. Stewart Ph.D.1

1Research Development & Dissemination, Sutter Health, Walnut Creek, CA

Introduction

Applications designed to improve care processes and that interface with electronic health record (EHR) systems are becoming increasingly common. But scalable approaches do not exist to systematically evaluate the effectiveness of these applications and, in particular, early stage process measures relevant to adoption and to workflow effectiveness. In our own experience there is a high bar for physician adoption and rapidly learning how to motivate adoption of new apps requires precise and timely data on whether, when, for whom, and how the app is being used.

CM-SHARE is a web-based application to improve primary care management of cardio-metabolic (CM) health by improving physician efficiency and by facilitating the closing of care gaps (i.e., improve evidence-based care) and by improving physician efficiency. CM-SHARE is designed and developed in collaboration with a team of primary care physicians. The application extracts real-time data from the EHR and other data sources and assembles them into an intuitive visual display of actionable information that the physician can access from the EHR during an encounter.

We describe our systematic approach in using web and audit log file data to measure adoption, use and the impact of CM-SHARE.

Methods

CM-SHARE was deployed in two primary care sites starting in April 2016. A total of six physicians were given access to the application through a link in the EHR. Physicians were provided training on the use of the application, after which subsequent use was completely voluntary. After go-live, we utilized a combination of qualitative and quantitative methods to understand adoption, use, and impact. We scheduled regular interactions with physicians to elicit qualitative feedback on their use of the application, its perceived value, and information on enhancements that would improve the usefulness of the application. We used two sources of audit log data to quantify adoption, use and impact: 1) web log data generated by the CM-SHARE application itself, which tracks the “who, what, where, when” of each user interaction with the application, and 2) EHR audit log data, an automatically generated, time-stamped activity log that tracks the “who, what, where and when” of each user’s interaction with the EHR.

Adoption was assessed by tracking the number of times CM-SHARE was launched, as well as the clinical and patient characteristics associated with each encounter by a participating physician from the EHR. To assess usage, we tracked which features of the application were used and not used, the duration of use, and the clinical and patient level correlates of use/non-use. To assess impact, we measured the clicks and time clinicians spent in the EHR when CM-SHARE was launched and compared these measures across similar patients/encounters prior to the go-live date, as well as in encounters where CM-SHARE was not used.

Results

Adoption: CM-SHARE was used consistently in 35-45% of encounters across the 16-month follow-up period (Figure 1), and in 50-56% of encounters where the reasons for the encounter was related to a cardiometabolic condition.

Use: Launching CM-SHARE is associated with advanced age, male sex, and an encounter visit associated with a CM reason, particularly diabetes. Qualitative feedback from participating physicians was used to improve/expand CM-SHARE functionality, including two major iterations: one which enhanced patient education features (version 2), and a second release (version 3) which reorganized the visualization to include key data points, e.g. a display of the patient’s medication adherence, and more prominent display of frequently used features.

Impact: We identified 3,847 patients who had an encounter with a participating physician, from whom 2,627 (68.2%) patients had at least one use of CM-SHARE during an encounter (intervention group) in the post-implementation period, thus 1,220 were potential controls. In the 12-month pre-implementation period, 3,236 eligible patients were identified. After patient-level matching, 1,037 encounters for 290 intervention patients, 1,008 encounters from 290 post-implementation matched controls and 2,901 encounters from 870 pre-implementation matched control patients.
Comparing intervention to controls, there was no difference on the overall encounter duration, regardless of encounter features (i.e. level of service). The CM app significantly reduced a provider’s time spent in the EHR while in the exam room by 25-35% (Figure 2) for diabetes- or hypertension-related encounters. Use of the application was also associated with a 21-36% reduction in the total number of EHR activities (i.e., “clicks”) in the EHR for diabetes and hypertension related encounters. The total number of a provider’s EHR activities was not significantly reduced.

Figure 1: CM-SHARE overall utilization rate after implementation

Figure 2: Summary of duration of clinic workflow and provider time with EHR for intervention and controls, stratified by level of service.

Discussion
Application-level and EHR audit log data provide valuable and quantifiable insights into the adoption, use and impact of a web-based application that is deployed in clinical practice. These data have identified the most valuable features of the CM-SHARE solution, and these insights have influenced how we iterate the application. By combining audit, clinical and demographic data, we have gained insight into the highest-value use cases for the app, i.e., patients with more severe CM conditions. Audit data have also allowed us to quantify the amount of time physicians spend interacting with the EHR. CM-SHARE use is associated with reductions in the time and click burden that physicians face with current-generation EHRs to access and visualize data. In future work, we will explore how the application can be adapted in real-time to the precise CM-related needs of an individual patient.

References

A Dynamic Analytic Process Description Framework for Big-Data
Boudewijn Aasman, BSc;1 Manuel Wahle, MS1; Saurabh Rawal, MS1; Shipak Banerjee, PhD1; Parsa Mirhaji, MD, PhD1
1 The Center for Health Data Innovations and Institute for Clinical and Translational Research at Einstein and Montefiore, Bronx, New York

Introduction

Following recent advancements in the realm of machine learning, analytics with healthcare data has seen a surge with multiple success stories. Despite the enormous potential, such analysis can be cumbersome due to data integration and data management challenges associated with big data (the 4 Vs of big-data). While data scientists spend the majority of their time to manage data instead of developing unique analytics, the costs and complexity of these processes mount a barrier for accessibility and pervasive use of big-data in healthcare.

Health analytics challenges

- Variability of source
- Variability of data structure
- Volume
  - Designing project specific ETL processes
  - Deployment of developed models

Big data challenges in healthcare

- Number of articles per theme (1)
  - Others
  - Lack of skills
  - Managerial issues
  - Storage & Standardization
  - Security
  - Data Structure

Another important part specified within the FDL is a formal description of all necessary data manipulation processes (e.g. ingestion, outlier removal, derived feature generation, and transformation) throughout the pipeline. FDL is extended to incorporate complex logic based computations typical of deep-phenotypes (formal definition of a cohort and relevant encounters), and to automate generation of registries for cohorts specified by users.

Target users of 'frames' generated by FDL include but are not limited to researchers, clinicians, data scientists, machine learning engineers, and automated agents, systems, services and pipelines.

Methods

The FDL uses ontologies to explicitly and formally describe concepts and functions to identify, access, and retrieve data using queries that are fully or partially formulated by automated agents, and instructions on how the data needs to be munged and staged before it is passed on to the downstream processes or user. Formalization and standardization of the frame description allows for separation of these aspects in such a way that the individual parts can be cataloged, invoked and reused.

FDL process: The box at the center depicts the FDL core. The code base is written in Python while all the transformations can be done with the option of Spark or Pandas and the scientific python stack. User defined functions or processes (e.g. an existing deep phenotyping algorithm or random forest classification model) can be incorporated into the FDL using an easy to use plugin system without any knowledge about the core code base.

FDL supported projects

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPROVE</td>
<td>Deployment</td>
<td>Machine learning model deployed as a Critical Care Decision Support System. EMR data is fed at 4 hour intervals via the FDL. Generated scores are used as a predictor for the danger of requirement of immediate mechanical ventilation and/or death.</td>
</tr>
<tr>
<td>Prediction of Resp. Failure (deep learning)</td>
<td>Research</td>
<td>A research project to improve the timeliness and sensitivity of time-sensitive critical analytics using LSTM and deep learning.</td>
</tr>
<tr>
<td>Severe Sepsis and Septic Shock determination</td>
<td>Audit</td>
<td>FDL feeds an analytical model used to determine the time of presentation of severe sepsis and septic shock among critical care patients. Runs on a quarterly basis to gather data for audit by the New York State.</td>
</tr>
</tbody>
</table>

Observations from implementation of the FDL

- Deploy complex analytic pipelines without coding
- Agnostic of data source (supports SQL (inplace), RDIMS and NoSQL)
- User has flexibility in choice of back-end
  - Apache Spark back-end is scalable and compatible with big data, while feature rich Pandas back-end can be good for one-off research projects
- Can support continuous deployment of analytics models via data streaming
- Can support continuous learning of machine learning models
- Flexibility in choice of front-end
  - Can be instantiated from multiple interfaces
- Maintain dynamic registries for each instantiation. This allows for reconstruction of the entire process in future for any purpose.

Conclusions

Several instances of the FDL have been used in Montefiore Health System to fully automate data pipelines for predictive models and analytics that utilize our underlying data lake infrastructure. They support several research projects, data delivery for audits, external reporting, and handle deployment of data based Critical Care Decision support systems. FDL instances are fully automated and have shifted the data scientists efforts from data management to data analytics.

Advantages of using the FDL

- Avoids analytic silos by returning and linking all analytical results and byproducts back to data.
- Allows integration of all data and analytics across all domains and applications for Audit, AI/ML, meta-analysis and patient empowerment.
- Saves the infrastructure and effort invested in creating ETL processes, designing schemas for various analytics and creating isolated analytics data marts.
- Allows for scalable machine learning pipelines through continuous learning and automation.

Bibliography


© 2016 MagiPrint Inc. www.publicdomain.com
A Subphenotyping NLP Algorithm for Identification of Critical Limb Ischemia Cases from Clinical Notes

Naveed Afzal, PhD; Vishnu Priya Mallipeddi, MBBS; Sungwan Sohn, PhD; Hongfang Liu, PhD; Rajeev Chaudhry MBBS, MPH; Christopher G. Scott, MS; Iftikhar J. Kullo, MD; Adelaide M. Arruda-Olson, MD, PhD; Mayo Clinic, Rochester, MN

Introduction

Lower extremity peripheral artery disease (PAD) affects millions of people worldwide. Advanced cases of PAD may manifest as critical limb ischemia (CLI), an adverse outcome associated with considerable morbidity, mortality and high risk of major cardiovascular events. Within one year of CLI diagnosis 30% of patients undergo limb amputation while 25% die. However, automated identification of cases is challenging due to absence of a single definitive International Classification of Diseases (ICD-9 or ICD-10) code for CLI. Prior studies have used CLI-related ICD-9 codes (440.22 - atherosclerosis of the extremities with rest pain; 440.23 - atherosclerosis of native arteries of the extremities with ulceration; 440.24 - atherosclerosis of native arteries of the extremities with gangrene) for identification of CLI cases. We compared the performance of the CLI-NLP algorithm with CLI-related billing codes. Both methods were compared to human abstraction as the gold standard. We tested the hypothesis that an NLP algorithm applied to narrative clinical notes will have superior performance compared to CLI-related billing codes for identification of CLI.

Methods

We applied the previously validated knowledge-driven NLP algorithm (PAD-NLP algorithm) to the dataset to automatically ascertain PAD status 1. The PAD-NLP algorithm automatically ascertained cases from clinical notes using PAD-related keywords and a set of rules for classification of a PAD patient. The PAD-NLP algorithm consisted of two main components: text-processing and patient classification. The text-processing component analyzed the text of each clinical note by breaking down sentences into words using MedTagger 2, an open source NLP system, and identified PAD-related concepts which were mapped to specific categories that were used for patient classification. CLI keywords were identified by cardiovascular experts and included in the list of keywords used in the PAD-NLP algorithm. A subphenotyping algorithm for identification of CLI cases was developed and used the document level output of the PAD-NLP algorithm and narrowed the focus of the algorithm to identify the subset of PAD cases with CLI.

Results

There were 295 CLI cases (37%) and 497 (63%) controls (without CLI). The average age of patients in the dataset was 71 years. The dataset was comprised of 44% women and 90% of patients were white.

<table>
<thead>
<tr>
<th></th>
<th>CLI-NLP (95% CI)</th>
<th>Billing codes (95% CI)</th>
<th>p-value CLI-NLP vs. Billing codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (%)</td>
<td>84 (79, 88)</td>
<td>88 (84, 92)</td>
<td>0.12</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>98 (97, 99)</td>
<td>74 (70, 78)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PPV (%)</td>
<td>96 (94, 98)</td>
<td>67 (62, 71)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>NPV (%)</td>
<td>91 (89, 93)</td>
<td>91 (88, 94)</td>
<td>0.87</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>93 (91, 95)</td>
<td>79 (76, 82)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Conclusions

The CLI-NLP algorithm for identification of CLI cases from narrative clinical notes had excellent PPV, specificity and accuracy. It has potential for translation to patient care, as it will enable automated identification of CLI cases for quality projects, clinical decision support tools and a learning healthcare system.

References

Predicting Pressure Injury among Surgical Critical Care Patients: A Machine Learning Approach

Jenny Alderden APRN, PhD CCRN, CCNS1,2, Mollie R. Cummins RN, PhD, FAAN2, Ginette A. Pepper RN, PhD, FAAN2, Andrew Wilson PhD, MSTAT2

1Boise State University, Boise, Idaho; 2University of Utah, Salt Lake City, Utah

Introduction

Hospital acquired pressure injuries (HAPI) are a common and serious problem among surgical critical care patients. Although pressure injuries are common, some pressure injuries (PIs) can be prevented using measures such as specialty beds that are not feasible for every patient due to cost1. However, decisions about which patient would benefit most from a specialty bed are problematic because existing pressure injury risk identification tools identify nearly all critical care patients as ‘high risk’2.

The purpose of our study was to develop a model that classifies pressure injury development among surgical critical care patients using machine learning methods. The model could be used to aid clinicians in identifying those at highest risk, for early intervention.

Methods

We used data from the electronic health record to develop a model to predict pressure injuries among surgical critical care patients at a level 1 trauma center. We performed our analysis in R version 3.3.2 via the RStudio interface. We utilized multiple imputation for missing data. We divided our data into training (67%) and testing (33%) datasets and induced a model using a random forest algorithm via the R package ‘randomforest’.

Results

The final sample consisted of 6,376 patients. In this sample, two hundred and eighty-three individuals (4.4%) developed HAPIs of category 2 or greater.

We induced a random forest (RF) model to predict category 2 and greater PI development among patients in the training data set, using cross-validation to avoid model overfitting. We used the testing data set to evaluate classifier performance. The area under the ROC Curve was 0.79 (95% CI 0.77-0.82). The five most important variables, according to the mean decrease in accuracy, were body mass index, surgery time, creatinine, hemoglobin, and age.

Discussion

One way to interpret our model’s performance is to compare it with that of the Braden Scale. The Braden Scale is the most commonly used PI risk-prediction tool in North America. Our model’s relatively strong performance (area under the ROC curve = 0.79 vs. 0.68 for the Braden Scale3) suggests it would be a useful way to differentiate among critical-care patients to apply preventive measures that are not feasible for every patient due to cost, such as specialty beds. The finding that surgical time was a relatively important variable in the analysis warrants further study as few studies have examined duration of surgery in relation to pressure injury risk.

References

Are You Ready to be a Single IRB? A Workflow for Electronic Review

Randall W Alexander, MS, Sherly Roy Yesudhas, MS, Gayathri Parthasarathy, MS, Stephanie Gentilin, MA, Brigitte White, MA, Stacey Goretzka, CIP, Summer Young, MPH, JD CIP, Susan Sonne, PharmD, Jihad S Obeid MD, Leslie A Lenert, MD MS

Medical University of South Carolina, Charleston, SC

Abstract

A Single IRB model can lessen the time needed for sites to receive IRB approval, but most academic institutions do not have IRB systems that address workflow issues associated with approving multiple sites within a single IRB application. A modification to an existing electronic IRB application allows processing of Single IRB reviews.

Introduction

Large clinical trials often require multiple sites in order to recruit participants in a timely manner. However, engaging multiple sites in one trial presents administrative burdens to obtain IRB approval at each site. The idea of a Single IRB (sIRB) model has the potential to lessen the time needed for a site to receive IRB approval, but most academic institutions do not have IRB systems that adequately address workflow issues associated with approving multiple sites within a single IRB application. The problem is that current electronic IRB applications do not support the sIRB process burdening the approving IRB and the lead study team. This poster describes a modification to our existing Huron Click electronic IRB (eIRB) application that allows processing of sIRB reviews lessening the burden on the approving IRB and the lead study team.

Methods

Our solution creates a separate Relying Site Documentation (RSD) study workspace for each sIRB multisite study. Here the relying site details are reported, reviewed, approved, and maintained. An RSD is similar to, but independent of study amendments. The RSD process increases efficiencies for the lead study team by including within the existing eIRB study application an indication that this is an sIRB study and identification of document requirements for the relying site, as well as RSD management on the relying sites’ behalf. The RSD helps the relying site by providing a secure, editable workspace for sites to upload customized documents, include study personnel, document reportable events and report recruitment data for continuing reviews. The RSD process benefits the sIRB reviewing the study by mirroring current IRB review processing within each RSD workspace.

Results and Discussion

Extending the Huron (Click) eIRB application to handle large multi-site trials provides the tools for the IRB and lead study team to effectively manage large multi-site studies. The extension provided the IRB and lead study team with a fully supported workflow for the entry, review, approval and documentation for the relying sites. The extension also allows the relying teams to assume the responsibility for entry of required documents and responding to reviews directly in the system. The flexibility in design allows the lead study team the option to serve as a proxy for the relying site by entering the required information on their behalf. MUSC began successfully using the design in March 2016.

Acknowledgements

This work was funded in part by the National Center for Advancing Translational Sciences of the National Institutes of Health, grant # UL1 TR001450 and The Duke Endowment through Health Sciences South Carolina.
Linking Clinical Events to Patient Provided mHealth Data

Nicholas Anderson, PhD\(^1\), Albert Riedl, MS\(^1\), Madan Dharmar, MBBS, PhD\(^1\), Meghana Gadgil, MD, MPH\(^2\), Michael Lehman\(^3\), David Lindeman, PhD\(^3\), Dan Gillette, EdM\(^2\), Brad Pollock, PhD, MPH\(^4\), David Haddad MSC\(^5\), Ida Sim, MD, PhD\(^6\)

\(^1\)University of California Davis, Sacramento, CA, \(^2\)University of California, San Francisco, CA, \(^3\)University of California Berkeley, CA, \(^4\)Overlap Health, San Francisco, CA

**Background** Integration of patient provided mobile health data into clinical health care environments from personal monitoring devices involves overcoming significant challenges of linkage, establishing data quality and provenance, and defining meaningful clinical utility. Legal issues of data ownership and participant privacy between consumer owned, clinical regulated and vendor systems further complicate requirements to allow patient and clinical systems to align and evaluate integrative data that meets both clinical and patient expectations.

**Objective** The PERCEPT (Personal and contextual precision health) platform provides an integrative, adaptable and participant-centric framework that links patient provided data to clinical events, and is designed to enhance patient engagement through providing common and shared views of data and progress between patients and health providers.

**Methods** PERCEPT (Figure 1) standardizes and links patient provided mHealth data captured through mobile phones with clinical data from their active EHR records, and protects and secures patients’ private personally generated mHealth data. Patients control all data they share through management on their own devices. PERCEPT connects these data sources via a secure deidentified IDs (1) provided by the study REDCap database (2) during patient enrollment. During monitoring phases, updates to patients clinical records are sent to a common i2b2-based data server (3), which monitors event changes relevant to individual patient monitoring plans (4). Actionable events are made available through a FHIR (Fast Health Interoperability Resources) interface to the study dashboard\(^5\) (5,6). mHealth data is monitored from enrolled patients mobile phone devices through commonly available consumer applications, and via as patient provided data on medication compliance and behavioral affect collected in the PERCEPT mobile application through a data integrator (7,8,9). Patients see the same data views, events, and clinical health data as their care team, viewing them on their personal phones vs. the study management dashboard (6).

**Discussion** The PERCEPT platform is engaged in a single-arm clinical observational trial of 200 hypertension and depression patients recruited across two medical centers and linked to the hosted environments. This pilot implementation is a novel approach to engaging and sustaining patient enthusiasm in their own health monitoring and medication compliance solely on their personally owned smart-phones. The evaluation of patients enrolled within this study are focusing on their sustained participation, comfort with privacy and security of the platform, ability to modify and communicate their health goals, and compliance with prescribed medications. This framework provides a low-burden approach to patients to be able to consistently provide access of their mHealth data and patient self-reported health status, and allow standardization of clinical events in a secure, shared and contextual information environment. The generalizability of this framework supports expansion to pragmatic engagement of large populations of chronic or acute disease management patients with limited technical or lifestyle change adoption requirements.

**References**

1. Personal data for the public good - New opportunities to enrich understanding of individual and population health. Health Data Exploration Project, Robert Wood Johnson Foundation, 2014
Informatics Infrastructure: Tip of the Spear for the Indiana University Precision Health Initiative

William K. Barnett, Ph.D.¹,²; David Pierce, BA²; Tony French, BS¹; Burke Mamlin, MD¹; Paul Biondich, MD, MS¹; Shaun Grannis, MD, MS, FAAFP¹; Peter J. Embi, MD, MS, FACP, FACMI¹,²; Tom Barber³; David Fuller³;
¹Regenstrief Institute, Indianapolis, IN; ²Indiana University School of Medicine, Indianapolis, IN; ³LifeOmic, Indianapolis, IN

Abstract
In 2016 Indiana University made a $120M investment in a Precision Health Initiative. This commitment requires a novel informatics infrastructure strategy with 3 conceptual components. A data commons is designed to capture and integrate phenome, genome, and exposome data. Applications driven by precision health use cases interact with the data through application programming interfaces. A customer relationship management platform provides metrics, improves services, and connects communities of patients, investigators, and care providers. We present the informatics infrastructure approach and progress to date.

Introduction
The goal of the Indiana University Precision Health Initiative (IUPHI) is: to position IU among the leading universities in understanding and optimizing the prevention, onset, treatment, progression and health outcomes of human diseases through a more precise definition of the genetic, developmental, behavioral and environmental factors that contribute to an individual’s health. This vision drives a three-component informatics infrastructure:

- A Data Commons that aggregates and phenome, genome, and exposome data. By developing a standards-based yet flexible data repository, we can collect and leverage the multi-factorial influencers of health outcomes and provide a data foundation for a precision health approach.
- A suite of applications that support the precision health activities of investigators, care providers, and patients. By implementing software-mediated precision health services, we can provide investigators the tools to better understand the dynamics of precision approaches to healthcare and to better identify and predict precision based approaches to care. For care providers, we can integrate those approaches into clinical decisions. For patients, we can provide tools for them to actively participate in the process of precision health.
- A customer relationship management platform that tracks activities and provides metrics and supports community engagement. By understanding the requirements of our research, clinical, and patient communities, we can respond to their needs, evaluate the success of our efforts, and ultimately help improve outcomes.

Methods
We leveraged active precision health clinical programs at the Indiana University School of Medicine. Stakeholder analysis presented us with numerous use cases to fulfill the mission of the IUPHI grand challenge. We collected over 200 initial user stories from 3 clinical teams (Triple Negative Breast Cancer, Pediatric Sarcomas, and Multiple Myeloma) that drove the initial development of: a Patient Viewer tool for clinical decision making; a Cohort Builder application for clinical research; general Visual Analytics tools for more flexible statistical analyses; Trial matching tools to rapidly discover appropriate clinical trials for patients; Patient Engagement Tools support volunteer registries, trial discovery tools for patients, eConsent, and patient reported outcomes and patient outreach tools.

Results
We are co-developing this infrastructure with our corporate partner LifeOmic, who provide a secure cloud precision health platform supporting researcher and clinical communities. LifeOmic’s Precision Medicine Platform™ helps manage, process, analyze, and share massive health data sets to improve and accelerate clinical decision making.

Discussion
The Indiana University vision of a precision health approach is driving the infrastructure conceptualization, design, and development. Key components of that design are a scalable Data Commons, flexible applications, and relationship management strategy across communities. A partnership among Indiana University, Regenstrief Institute, and LifeOmic is a promising modular, secure, and scalable model for precision health.

Acknowledgements
This project is funded through the Indiana University Grand Challenges program, https://grandchallenges.iu.edu.
Mapping of HIE CT codes to LOINC - Analysis of Inconclusive Codes and Quantification of Mapping Times

Anton O. Beitia, MD¹, Tina Lowry MS¹, Daniel J. Vreeman, PT, DPT, MS², Bradley N. Delman, MD¹, George Loo, DrPH¹, Frederick Thum III, MD¹, Jason S. Shapiro, MD¹
¹Icahn School of Medicine at Mount Sinai, New York, NY; ²Regenstrief Institute, Indiana University School of Medicine, Indianapolis, IN

Introduction
Mapping proprietary computed tomography (CT) terms to a standard vocabulary is necessary for interoperability and building real-time decision alerts. This study describes key characteristics of the mapping process.

Methods
We mapped CT exam codes in our local health information exchange (HIE) to LOINC terms using LOINC’s mapping tool, RELMA®. From a total of 10552 mapped HIE CT codes, we tracked straightforward codes (mappable to LOINC based solely on the provided exam names) and inconclusive codes (not mappable without further investigation) and calculated the prevalence of each code type. We also analyzed semantic and syntactic differences in exam names between randomly selected 100 straightforward and 100 inconclusive CT codes. We also randomly chose 30 straightforward and 30 inconclusive codes and measured the time to map each code to a LOINC or “temporary” (instances without LOINC match) code. Mean times were calculated for all codes and both sets.

Results
The majority of codes were straightforward (73.5%) and a minority were inconclusive (26.5%). Straightforward exam names typically contained unambiguous specification of imaging modality, anatomic region, presence of contrast, and timing and route of contrast administration (e.g., CT neck WO&W contrast IV). For procedural codes, specific action (e.g., biopsy) and anatomic object (e.g., liver) are also associated with straightforward naming.

Inconclusive exam names were retrospectively grouped into categories, which are listed with relative frequencies in Table 1. The categories are non-mutually exclusive, so the number of exams totals >100.

<table>
<thead>
<tr>
<th>Inconclusive exam name category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unambiguous exam name without current LOINC match, but distinguishable by attributes such that a LOINC term would likely be created if requested (e.g., CT guided biopsy left kidney)</td>
<td>37</td>
</tr>
<tr>
<td>Exam name lacking enough detail to map without further investigation (e.g., whether local code CT Abd/Pel AAA maps to CT Abd/Pel W IV or CT Abd/Pel WO&amp;W IV)</td>
<td>29</td>
</tr>
<tr>
<td>Anatomic focus in exam descriptive name is non-specific or ambiguous (e.g., CT orbit/sella/ear)</td>
<td>27</td>
</tr>
<tr>
<td>Exam name contains term with no meaning outside a specific facility (e.g., Code T Head CT)</td>
<td>7</td>
</tr>
<tr>
<td>Exams with name attributes outside the LOINC model scope; including consultative (e.g., CT outside read) and administrative type (e.g., CT ambulatory) exam names</td>
<td>7</td>
</tr>
<tr>
<td>Exam name includes a highly specific reason for exam (e.g., CTA Chest/Abd/Pel for TAVR)</td>
<td>6</td>
</tr>
<tr>
<td>Exam name contains multiple imaging guided procedures grouped together under one code (e.g., CT guided needle placement, biopsy, or aspiration)</td>
<td>2</td>
</tr>
<tr>
<td>Exam name contains a non-standard imaging location (e.g., CT Left Abd/Pel).</td>
<td>2</td>
</tr>
</tbody>
</table>

The mean ± S.D. mapping times for straightforward and inconclusive codes were 0.48 ± 0.15 and 9.55 ± 4.62 minutes respectively, a twentyfold difference that was statistically significant (p<.001, Wilcoxon rank sum test). Average mapping time/code for all code types was 2.89 minutes, calculated by multiplying mean times for mapping straightforward and inconclusive codes by their prevalence rates respectively.

Discussion/Conclusion
We identified distinguishing characteristics of inconclusive CT codes that preclude mapping to LOINC on the basis of exam descriptions alone. Although the prevalence of these codes is lower than straightforward codes, they require significantly more time and effort to map. By noting the contrasting characteristics of straightforward and inconclusive codes, our work may assist others in local code description development for improved interoperability.
Using Knowledge from Quality Measures for Predicting Mortality and Survival in Patients with Acute Myocardial Infarction

John W. Bisognano, Aditi Gupta, PhD, Matthew Lui, Albert M. Lai, PhD, Philip R.O. Payne, PhD
Institute for Informatics, Washington University in St. Louis, MO, USA

Introduction

Most disease-outcome association analyses, use GWAS (genome-wide association studies) and PhêWAS (phenome-wide association studies) based approaches to draw from directly recorded EHR or genomic data. We propose to augment these approaches with measure-based features developed using the canonical knowledge from known quality factors and clinical guidelines to improve the predictive power of association studies. Quality measures are metrics used to quantify and evaluate healthcare delivery and are reported as claims, charts, and registries. Because most information needed to calculate quality measures is already recorded, it is possible to build approximations of these quality measures from EHR data. National Quality Forum (NQF) and National Quality Measures Clearinghouse (NQMC) provide text/XML format descriptions of quality measures.¹ Traditionally, these quality factors have been used as ambulatory measures to access the quality of care, but our goal is to use the hidden canonical knowledge and relationships between these measures and the disease, to predict the outcomes. We demonstrate the effectiveness of our approach for Acute Myocardial Infarction (AMI) which has about 790,000 cases each year in the United States.

Methodology

We applied a three-step methodology for our approach based on utilizing knowledge extracted from quality factors for identifying and evaluating associations in AMI patients' cohort. Step 1) We mapped/coded the information from the quality measures into a combination of EHR derived measures using ICD9-codes, procedures codes, and clinical notes based extraction. For our preliminary analysis, we selected three different quality factors related to AMI based on Body Mass Index (BMI) (NQF# 2828), tobacco use (NQMC# 010141) and blood pressure (NQF# 0018), we focus on using the knowledge and relationship in these measures; Step 2) We computed the values for the quality measures from the EHR dataset based on the mapping created in Step 1. Analysis was done using MIMIC-III (Medical Information Mart for Intensive Care) [1], which is a freely-accessible, large single-center database containing deidentified patient data for patients admitted to critical care units to Beth Israel Deaconess Medical Center (Boston, MA); Step 3) We evaluated the associative ability of these measures by viewing their correlation with outcomes (60-day survival and 60-day mortality) in patients presenting to the hospital with acute myocardial infarction. For this step we applied linear regression, to draw associations between the quantified quality factors and our outcomes.

Results

We present preliminary results for one of the above-mentioned quality factors that we analyzed: "2828 - Preventive Care and Screening: BMI Screening and Follow-Up Plan". Applying the NQF factor to the MIMIC population resulted in the creation of two subgroups based on the quality factor's criteria: one subgroup of 2,238 patients between the ages of 18-65 and 2,244 patients older than 65 years old. Using linear regression, we found that the BMI quality factor was statistically significant ($p$-value < 0.01) in predicting the survival or readmission outcome for the subgroup of patients older than 65 years old, but for the subgroup of patients ages 18-65, we did not find strong associations.

Discussion and Conclusion

In this work, we presented our approach of using measure-based features developed using the knowledge extracted from quality factors for predicting disease outcomes. Our preliminary results indicate that we can successfully utilize quality measures for predicting clinically important outcomes for patients with AMI. Some of these NQF quality measures, such as the tobacco and blood pressure factors, could not be coded in full as some information was not readily available and/or reliably recorded in the EHR. Further analysis could apply this measure-based analysis to hospital data to evaluate the broader impact of these quality measures.

References


¹ NQMC: https://www.qualitymeasures.ahrq.gov ; NQF: https://www.qualityforum.org/
Achieving New Results with What Already Exists:
Repurposing an existing standard and functionality to achieve new goals

Cory M Blackwell, BS\textsuperscript{1}, Kathleen Palso, MA\textsuperscript{1}, Alexander Marfo, BS\textsuperscript{1}, Kristi R. Eckerson, MSPH\textsuperscript{2}, Brian Gugerty, DNS, RN\textsuperscript{1}

\textsuperscript{1}National Center for Health Statistics, Hyattsville, Maryland; \textsuperscript{2}Karna, LLC, Atlanta, Georgia

Problem Addressed and Specific Purpose of the System

The National Center for Health Statistics (NCHS) is continually seeking ways to boost quality and participation in its surveys of nationally representative samples of health care providers. Various legislative initiatives have made electronic health care records (EHRs) particularly attractive as a method for data collection, namely the HITECH Act included in the 2009 federal stimulus package. NCHS saw opportunity in the increasing support of EHR interoperability and conducted pilot studies to explore the feasibility of collecting data using Continuity of Care Documents (CCDs). NCHS also developed the HL7 CDA® R2 Implementation Guide: National Health Care Surveys Release 1, DSTU (IG) as a standard for the submission of EHR data for its health care surveys. In late 2015, the National Health Care Surveys were named as one of the measures by which providers could attest to meeting one of the public health reporting objectives of the EHR Incentive Programs (Meaningful Use or MU). The timing of this inclusion created an urgency to register physicians for the program and to develop a way to collect data from sampled physicians electronically. The problem was that EHR vendors had not yet implemented the IG. Taking advantage of existing Consolidated Clinical Document Architecture (C-CDA) standards and implementations, NCHS designed a system to collect CCDs. While the CCD standard is a single C-CDA document standard, each vendor interprets that standard slightly differently, such that each implementation of an EHR may exhibit some variation from the others. These idiosyncrasies manifest in the “standard” documents produced by each system. Despite issues with processing documents that vary, NCHS has been able to successfully extract encounter-based survey data from CCDs for its health care surveys.

Evaluation of Results

At the time of sampling for our 2016 National Ambulatory Medical Care Survey (NAMCS), approximately 131,000 Eligible Providers (EPs) had registered as willing to provide survey data under the MU program. Of those EPs, 537 were in the NAMCS sample and invited to participate in the survey. Application of final eligibility criteria resulted in 336 being invited to Testing and Validation (T&V), during which 2,140 example CCDs were submitted. During validation, no CCD was found to be error-free. The number of errors and warnings per file ranged from 19 to 1,161 (median = 94). All sampled EPs were invited to Production. To date (09/13/2017), NCHS received 4,677 CCDs containing production survey data resulting in 4,221 usable encounter records.

Conclusion

The current NCHS-designed system addresses difficulties with document variability as it processes CCDs into usable data for the 2016 data year of the NAMCS. The coded entries (as opposed to content intended to be presented for the human eye) within the structured documents were often lacking in adherence to the standard; however, despite some processing issues, collection of encounter data from CCDs was shown to provide a new way to obtain public health data, and CCDs are a viable alternative to the labor-intensive medical chart abstraction by field representatives that NCHS had previously relied on.
Computational Cost and Complexity of Implausible Value Detection Methods

John P. Borsi, BS¹, Wei Yao, MS¹, Nnenna Ibeanusi, MSPH/MSIS¹
¹IBM Watson Health, Cleveland, OH

Introduction

Lawman et al¹ reviewed different methods for identifying Biologically Implausible Values (BIVs) used by epidemiological studies. They identified several different approaches for identifying BIVs based on longitudinal methods, including Kim et al², Lawman et al³, and Sturm & Datar⁴. Additionally, some methodologies have been published without an accompanying epidemiological study, such as Muthalagu et al⁵ and Daymont et al⁶.

Although Lawman, et al’s work compared the results of various methodologies on data retention and aggregate statistics, no work has examined their computational characteristics. Those considerations, such as code complexity and runtime, are essential to scalable and replicable results on extremely large datasets.

Methods & Results

We implemented the BIV detection methods cited above in a custom-built modular Java framework within the Hadoop MapReduce paradigm. The code was run over 55 million patients with 600 million height and weight records in the IBM Explorys Therapeutic Dataset, which is derived from clinical and claims records from several large US hospital systems. This processing occurred in a 243 node cluster. We compared the running times, code length, code complexity, and data results of the implemented algorithms. The results of our tests are given in Table 1.

Table 1: Results of Implementation Tests for BIV algorithms.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Code Size (lines)</th>
<th>Cyclomatic Complexity</th>
<th>Runtime (min/1M records)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim⁵</td>
<td>66</td>
<td>42</td>
<td>0.13</td>
</tr>
<tr>
<td>Lawman³</td>
<td>33</td>
<td>20</td>
<td>0.17</td>
</tr>
<tr>
<td>Sturm⁴</td>
<td>23</td>
<td>13</td>
<td>0.29</td>
</tr>
<tr>
<td>Muthalagu⁵</td>
<td>92</td>
<td>63</td>
<td>0.20</td>
</tr>
<tr>
<td>Daymont⁶</td>
<td>310</td>
<td>196</td>
<td>2.07</td>
</tr>
</tbody>
</table>

Conclusion

It is essential informed choices are made about the algorithms used to filter implausible values. As informatics are deployed on increasingly large stores of medical records, the timeliness, reliability, and usefulness of data will depend on our ability to manage these concerns. The feasibility of algorithms should be a primary concern for future work.

References

Patients concerns about the AP-HP Clinical Data Repository

Stéphane Bréant¹, Claire Hassen-Khodja¹, Armande François¹, Marie-France Mamzer¹, Gilles Chatellier¹, Christel Daniel¹,²

¹ Assistance Publique – Hôpitaux de Paris (AP-HP), Paris, FRANCE ; ² LIMICS, UMRS_1142 INSERM, Paris, FRANCE

Abstract

As required by the French law on data protection, Assistance Publique – Hôpitaux de Paris (AP-HP) patients were informed about the purposes of the AP-HP Clinical Data Repository and their right to object to the processing of their personal data for biomedical research. This poster presents the effects of patients’ individual information and demonstrates that ethical issues need to be addressed effectively at the early phase of the project to fully realize the potential of big data in medical research.

Context: the AP-HP Clinical Data Repository (CDR) (8M patients) consolidates, on a daily basis, clinical data extracted from the hospital information system to support observational studies, comparative effectiveness and clinical research. The AP-HP data warehouse is based on a state-of-the art IT infrastructure with PostgreSQL/HDFS storage and computing clusters. Data is accessible through i2b2 and Jupyter frameworks and allows the re-use of clinical data for research purposes for patients who did not opt-out. This infrastructure has been authorized by the French data protection authority in 2017. AP-HP has made a compliance declaration for studies which use existing data already collected for clinical care and for which the patient does not object to participate after having been informed individually about the use of their deidentified data in IRB-authorized studies and their right to opt-out. A major challenge is to get the fair balance between supporting medical research - by increasing efficient access to data stored in research platforms - and ensuring the effectiveness of data subjects’ rights to confidentiality, transparency and to opt out to the sharing of their data.

Objective: to describe the impact of the patient individual information campaign launched by AP-HP, as required by the French law on data protection, about the existence and purposes of the AP-HP CDR and their right to object at any time, without justification, to the reutilization of their personal data for the purpose of biomedical research.

Methods: On July 31, emails were sent to 476,145 patients with valid email addresses extracted from the AP-HP clinical information system. On August 29, paper mails were sent to 13,015 additional patients, selected at random. Patient information included the description of the AP-HP CDR, the opt-out procedure and the link to Frequently Asked Questions (FAQ) page of AP-HP clinical research and innovation office. A procedure was organized to address efficiently opt-out requests and patients’ concerns.

Results: as results of this information campaign, the website received more than 8,000 visits on July 31 (usually 300 visits per day), 634 patients opted out of processing their individual data for biomedical research whereas only 7 had done so within 3 months before the individual information campaign. Moreover 65 patients and 2 patients associations reached the institution to express 99 sensitive concerns about patient consenting (opt-out procedure (33%), opt-in (18%), patient information (16%), privacy and security (12%), patient engagement in research activities (11%), patient right to access their data (7%) and data property/donation and patient remuneration for data reuse (2%)). The institution responded to the patients and the FAQ page was updated accordingly for patients’ better understanding of AP-HP CDR and for patients’ reassurance about trustworthy reuses of their personal data.

Conclusion: Individual information of patients about their rights has a major impact on the number of opt-outs. Ethical issues need to be addressed effectively at the early phase of the project, and individual information content needs to be improved, to fully realize the potential of big data in medical research.
MPIO - A novel Minimum Potential drug-drug interaction Information Ontology implemented in OWL

Mathias Brochhausen, PhD¹, Maria Herrero-Zazo, PhD², Elizabeth A. Garcia, BA³, Daniel C. Malone, RPh, PhD⁴, Richard D. Boyce, PhD⁵

¹University of Arkansas for Medical Sciences, Little Rock, AR; ²King’s College London, London, UK; ³University of Pittsburgh, Pittsburgh, PA; ⁴University of Arizona, Tucson, AZ

Introduction

New evidence regarding potential drug-drug interactions (PDDIs) is published daily in primary sources such as drug product labeling and the scientific literature. Currently, no minimum information model standard exists to guide persons who synthesize PDDI evidence on how to organize the information in a way that would be most effective for use in clinical settings. In addition, current ontologies in the drug interaction domain, such as DINTO and DIDEO, are based on different conceptual models, as described by Herrero-Zazo et al. In order to fill these gaps, a public task force was created that operates within the World Wide Web Consortium Healthcare and Life Sciences Interest Group (http://www.w3.org/wiki/HCLSIG/DDI) and has involved more than two dozen international participants including pharmacists, physicians, and informaticists. In this poster we present a new ontology called the Minimum Potential drug-drug interaction Information Ontology (MPIO), a key component of the new standard being developed by the task force. We are in the process of applying the ontology to formalizing more than 10 detailed PDDI descriptions.

Methodology

The task force agreed on nine core PDDI information items based on recommendations from a prior conference series involving a wide range of stakeholders: evidence, mechanism, recommended action, frequency of exposure, frequency of harm, medical context, clinical consequences, seriousness, and severity. A user-centered definition for each term was created through an iterative process involving all task force members. After the task force agreed upon on the scope of knowledge representation for the new standard, two ontology experts (MB, MH) created a genus-differentia definition for each user-centered definition and started implementation in OWL2.

Results

The MPIO is disseminated through a GitHub repository (http://github.com/MPIO-Developers/MPIO) and can be accessed from (http://purl.obolibrary.org/obo/mpio.owl). MPIO provides a concise and semantically-rich OWL representation of the nine information items. At the time of this writing, 3 of 9 core information items are represented with formal axioms. A full list of terms and definition in MPIO can be accessed from: (http://owl2tl.com/e2d94210). The MPIO OWL file contains only 37 classes and acts as a ‘thin’ representation that leaves out representational overhead from upper ontologies. However, MPIO can be merged with much more detailed drug-related ontologies, such as DINTO and DIDEO, using shared classes from other OBO Foundry ontologies.

Conclusions

MPIO provides formal definitions and URIs for each of the core PDDI information items making it a key component of the new PDDI minimum information model standard. Moreover, it fills a gap if drug knowledge representation by representing nine core information items that are related but not fully within the scope of DIDEO or DINTO.

References

Development of a Sickle Cell Computable Phenotype for Predicting Disease Characteristics Using An Observational Medical Outcome Partnership Common Data Model

H. Timothy Bunnell, PhD, Erin L. Crowgey, PhD, E. Anders Kolb, MD
Nemours Alfred I. duPont Hospital for Children, Wilmington, DE

Introduction
Historically, many sickle cell disease (SCD) patients who manifest serious related adverse events (e.g. stroke and acute chest syndrome) will experience a severe course and die before the third decade of life. Some recent interventions, such as the use of hydroxyurea, do prolong life and reduce risk of SCD-associated adverse events. Other, more intensive therapies may also be applied. Unfortunately, these therapies provide variable benefit, and the treatments can be quite toxic. Thus, identifying patients early in life who are at greatest risk for severe disease, and who may not benefit from long term toxic therapy, is critical for improved care and outcomes. Toward this end, we are interested in (a) developing computable phenotypes (CP) able to accurately identify SCD patients in large EHR datasets such as PEDSnet [1], and (b) showing that it is feasible to use EHR data to isolate meaningful sub-phenotypes among SCD patients. These are seen as necessary preliminaries to the application of Learning Health System principals in improving care and outcomes for SCD patients.

Methods
Nemours maintains a research-ready extract of our EHR data in the PEDSnet variant of the OMOP Common Data Model (CDM). These data were used in two studies to (a) evaluate a CP based on the PEDSnet CDM, and (b) demonstrate classification of SCD patients by genotype. For the CP, we trained randomForest (RF) classifiers [2] and compared performance to a previously published SCD CP [3]. For comparability with [3], we used only ICD-9 codes in the RF analysis. The features for RF analysis were the relative frequency of each of 10 selected ICD-9 codes in each patient’s history (N=505).

In the second analysis, we classified subsets of SCD patients based on their disease genotype (HbS, HbC, Beta-plus and Beta-0; hereafter ss, sc, sb+ and sb0 respectively). These genotypes are known to be associated with differential risk for adverse events. We selected 20 features based on the contrast in the relative frequency of items in the CDM fact tables among SCD patients and matched non-SCD patients. Linear discriminant (LD) classifiers [4] were trained on these to identify the four genotypes. Leave-one-out cross validation was used to estimate LD accuracy and a manual chart review was conducted and used as the gold standard to benchmark the analysis.

Results
Table 1 compares results of the RF-based CP to those of [3] in terms of the number of true positive (TP), false positive (FP) and false negative (FN) cases as well as the associated positive predictive value (PPV) and sensitivity (SNS). The RF approach provided more accurate results in all measures. Table 2 shows the cross-validation results of the LD analysis of SCD genotype. The LD functions correctly classified 96.4% of the held out patient genotypes. 

Discussion
These preliminary analyses illustrate the potential power of a data-driven approach to analyses of EHR data. One limitation of the present approach was that due to the relative small number of training cases available, it was necessary to use artificial methods of estimating behavior on unseen data. In ongoing work we are validating this approach using an additional roughly 600 patients for whom hand curated information will soon be available.

References
Novel REDCap Utilization for a Clinical Study with Repeating Events

Ravan Carter¹ and Timothy D. Imler¹,²

¹. Regenstrief Institute, LLC. Indianapolis, IN.
². Division of Gastroenterology and Hepatology; Indiana University School of Medicine. Indianapolis, IN.

DESCRIPTION OF PROBLEM

Optimizing Patient Transfers, Impacting Medical quality, and Improving Symptoms: Transforming Institutional Care (OPTIMISTIC), is a demonstration project that was created as part of a Centers for Medicare and Medicaid Services initiative in September 2012 at Indiana University. The primary goal of the project was to reduce avoidable hospital transfers among long-stay nursing home residents.

Data was captured utilizing a HIPAA compliant REDCap¹ instance from both clinical and nursing home records. Early design set each row being a resident with an “event” as a resident’s stay in the facility. Each event could have up to 150 individual columns for variable capture and this repeated across the same row for each nursing home stay on an individual patient. This led to an unwieldy back-end data set that had thousands of columns that often were null as REDCap does not support a relational database structure.

The data capture process was then changed, in October of 2016 (Figure 1), to allow each row to have a unique identification for a resident, a nursing home stay, as well as a “block” of 10 data elements for the columns. This process allowed for a standardized method for identification of what data was contained within each row/column without extending the column length resulting in multiple null entries.

A back-end transformation (ETL) was then utilized to withdraw the data from the REDCap instance and translate it across the standardized row/column into a MySQL database that could then be used for research analysis (Figure 2). The ETL reads a “Map” (data dictionary) and takes each variable from REDCap and writes this to the MySQL database. Each night the entire MySQL database is dropped and the ETL process is run again to recreate the table and data structure. This recreation is due to the REDCap’s inability to be transactional in nature.

LESSON LEARNED

REDcap is a powerful and secure tool for capturing clinical data for research that is seeing increasing use throughout the academic community.

While recent updates to REDCap have allowed for repeating rows to be more easily addressed for complex data sets, there is still a pressing need to improve the back-end data storage structure to support relational data to simplify the process for research analysis without complex extraction procedures.


This abstract was supported by NIH/NCRR Colorado CTSI Grant Number UL1 RR025780. Its contents are the authors’ sole responsibility and do not necessarily represent official NIH views.
Towards the Implementation of a Canadian Pediatrics Diabetes Registry

Nelson Chan, BSc\textsuperscript{2}, Shazhan Amed, MD, MSc\textsuperscript{1,2}, Daniel Metzger, MD\textsuperscript{1,2}, Elizabeth Rosolowsky, MD\textsuperscript{3,4}, Rose Yeung, MD\textsuperscript{5}, Lawrence Richer, MD, MSc\textsuperscript{3}, Elodie Portales-Casamar, PhD\textsuperscript{1,2}

\textsuperscript{1}Department of Pediatrics, University of British Columbia, Vancouver, BC, Canada; \textsuperscript{2}BC Children’s Hospital, Vancouver, BC, Canada; \textsuperscript{3}Department of Pediatrics, University of Alberta, Edmonton, AB, Canada; \textsuperscript{4}Stollery Children’s Hospital, Edmonton, AB, Canada; \textsuperscript{5}Department of Medicine, University of Alberta, Edmonton, AB, Canada.

Introduction

Diabetes is a common pediatric chronic disease that is associated with serious long-term complications. Administrative population-level datasets such as physician billing, hospital admissions, and prescription dispensations have shown that there is a wide gap between treatment recommendations and the actual delivery of pediatric diabetes care. New interventions are necessary and a Canadian pediatric diabetes registry is essential to effectively evaluate the impact of these interventions and to improve the quality of pediatric diabetes care and patient outcomes across Canada. While relatively easy to handle at a local level, Disease registries can become difficult to implement nationally, with many possible strategies and infrastructures for data collection and sharing, resulting in sometimes-costly highly-tailored solutions that cannot be re-utilized. Moreover, these registries are very rarely integrated with the clinical systems, with data being collected in isolation and duplication.

Rationale

Our project aims at implementing a Canadian Pediatrics Diabetes Registry that draws data from existing clinical systems available at pediatric diabetes clinics across the country and enables benchmarking of current clinical practices and patient outcomes across institutions and jurisdictions. Based on availability (open-source) and user-base (>100 US hospitals, as well as internationally), we chose to implement the NIH-funded Informatics Integrating Biology & the Bedside (i2b2) data warehousing framework (https://www.i2b2.org). The modeled architecture includes a separate i2b2 warehouse for each Canadian province and to connect them together through a SHRINE data sharing network (https://www.i2b2.org/work/shrine.html). As a proof-of-concept, we decided to focus on two provinces (Alberta and British Columbia) and start implementation with the institutions with the largest flow of pediatric diabetes patients, BC Children’s Hospital (BCCH) and Stollery Children’s Hospital (Stollery).

Results

A critical step in the implementation of a national registry is to define a common data model to map the data coming from the disparate systems. We took a consensus-based approach to review the various data models in use and existing standards, and agreed to adopt the standardized minimum dataset defined by the SWEET initiative (http://www.sweet-project.org/downloads/SWEET-Dataset-Explanation.pdf). We mapped the data model in use in the custom-tailored database at the BCCH pediatrics diabetes clinic to the SWEET minimum dataset and rolled-out the BC i2b2 infrastructure. Efforts are still underway at Stollery where the clinical system in place is EPIC and no tailored data collection has been implemented for pediatric diabetes, rendering the data more difficult to map to the standardized minimum dataset. We are currently moving to the next phase in BC which involves expanding our data sources to additional pediatric diabetes clinics across the province. We are actively recruiting BC patients into the registry and consenting them for their data to be linked to other datasets beyond the registry to facilitate future research.

Conclusion

This initiative, to our knowledge, is a first effort in Canada to step beyond the world of research and test the utility of such integrated framework to interrogate data currently harboured in disparate clinical systems. The challenges in such an implementation is that it assumes that the necessary data are available in the clinical systems in the first place. Unfortunately, there is still a lot of the critical information that is embedded in unstructured clinical notes and is not readily available. In such cases, we will partner with the clinics to provide appropriate tools for structured data collection (e.g. REDCap). For the success of the endeavor, it is also essential to onboard both clinician and informatics champions at each institution to help navigate the local regulations and processes. This work is supported by the Maternal, Infant, Child, and Youth Research Network (MICYRN) and BCCH Research Institute.

496
Predicting Postoperative Bleeding Risk for Patients Undergoing Colorectal Surgery Using Machine Learning

David Chen, PhD¹, Naveed Afzal, PhD¹, Sunghwan Sohn, PhD¹, Elizabeth B, Habermann, PhD², James M. Naessens, ScD.², David W. Larson, MD, MBA³, Hongfang Liu, PhD¹

¹Division of Biomedical Statistics and Informatics, ²Division of Health Care Policy and Research, ³Division of Colorectal Surgery, Mayo Clinic, Rochester, MN

Introduction
Postoperative bleeding (POB) is a fairly rare but serious complication following colorectal surgery (CRS), occurring in between 2-14% in all CRS cases¹². There is a lack of consensus for risk factors for POB, suggesting the complexity of the underlying cause and lack of standardization of practice. Machine learning (ML) has been touted as a way to learn models that represent difficult non-linear systems, particularly when utilizing high-dimensional “big data”. The objective of this study is to present a predictive ML model for identifying patients at high risk of POB using data extracted from electronic health records (EHR).

Methods
A retrospective review of adult patients who underwent CRS as the primary procedure was conducted at Mayo Clinic Methodist Hospital between January 1998 and December 2014. POB was determined through ICD-9 code for any type of hemorrhage. A total of 13,399 CRS cases were found, with 1,680 cases (12.5%) having experienced POB. Ninety three variables falling under categories of demographics, patient provided information, symptoms, comorbidities, physiological values, observational factors, and terms extracted from clinical notes were extracted from EHR. Terms extracted from EHR were generated using MedTagger, an open-source tool which enables discretization of data that are locked in unstructured clinical notes.

Logistic regression (LR), and gradient boosting machine (GBM) methods were used to predict POB. The Boruta method, an iterative method which compares variable importance with shadow variables, was used for variable selection. Evaluation of each model was performed using 10-fold cross-validation. Due to the imbalanced class distribution, the observations with positive POB of the training set were oversampled while negative observations were undersampled such that each training set had a near 50/50 balance. All models were trained and evaluated in R.

Results and Discussion
Table 1: Performance of various models. ‘All’ refers to all pre-operative and intra-operative variables. ‘Known’ refers to well agreed upon risk factors (e.g. bleeding disease, comorbidities, surgical length, etc.). P-value (in parenthesis) is with reference to all variables.

<table>
<thead>
<tr>
<th>Subset</th>
<th>All</th>
<th>Boruta</th>
<th>Known</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.720 (1.00)</td>
<td>0.735 (0.35)</td>
<td>0.672 (&lt;0.01)</td>
</tr>
<tr>
<td>GBM</td>
<td>0.819 (1.00)</td>
<td>0.822 (0.72)</td>
<td>0.752 (&lt;0.01)</td>
</tr>
</tbody>
</table>

Table 1 shows performance of various different models. Surveillance for POB is a resource intensive task due to need for human attention and laboratory testing. Early and targeted surveillance can improve resource allocation, particularly if early intervention can reduce the amount of transfused blood. ML methods can often do a better job of modeling these systems because the models do not assume linearity. The results presented in this work demonstrate the utility of ML methods in predicting POB following CRS procedures. New risk factors such as mobility issues, weakness, nutrition level, and fall risk score were identified which could lead to improved understanding of causes of POB and to better allocate resources for POB surveillance.

References
Patient Compliance Documentation using REDCap, Twilio, and Video Uploads from Mobile Devices

John Clark, Erin Quigley, Aimee McRae-Clark, PharmD, Amanda M. Wagner, LPC, Jihad S. Obeid, MD, Leslie A. Lenert, MD, MS

Biomedical informatics Center, Medical University of South Carolina, Charleston SC USA

Abstract

Documentation of patient compliance with therapy in clinical studies is always difficult. Directly observed therapy is a gold standard approach for documentation. We have created a tool kit for documentation of compliance via text message prompted video capture of medication consumption.

Introduction

Documentation of compliance with medication prescriptions is critical aspect of pharmaceutical research. Various approach have been used to “prove” that patients have been compliant with the medication under study. These include pill counts, use of blister packs for medication, and electronic pill bottle caps that record openings. With the exception of biomarker tests, which are applicable only to a small number of pharmaceuticals, most measures of compliance stop short of certain knowledge that the prescribed medication has been taken. For example, pill bottle openings do not necessarily lead to drug ingestion. Therefore, the gold standard for measurement of adherence has been Directly Observed Therapy (DoT). DoT is difficult to implement and inconvenient for patients. We describe an approach for DoT implemented within REDCap using short message service (SMS) texts for video confirmation via mobile devices.

Methods and implementation

A pharmacotherapy clinical study was set up in REDCap using the longitudinal model with daily encounters (Figure 1). REDCap Automated Survey Invitations were setup to send SMS text with a survey link to participants twice per day using Twilio service (Figure 2). Consented participants are instructed to document medication ingestion using small video clips captured on their mobile device when prompted via the SMS link and upload videos via the survey. Videos are downloaded using the REDCap API with custom PHP code. Video metadata including the timestamp is extracted using ExifTool1 and inserted into the participant record. The code also checks the timestamp to ensure the uploaded video is within the specified time limit to take medication and not from an earlier time.

Results

The system was piloted on 22 participants, with a total of 1135 text messages, which prompted 862 video uploads through mobile devices, yielding a 76% overall validation rate.

Discussion

DoT was feasible to implement using text reminders and video uploads. Previous medication trials conducted in the same clinical population as the present study have reported adherence rates as low as less than 20% using biological markers as indicators of compliance2 and poor concordance between objective measures of compliance and participant self-report3. This methodology could be applied to other clinical trials or research studies where ingestion of medications or other activity needs to be observed by investigators as part of the consented research.

Acknowledgments: This work was supported in part by NIH/NCATS Grant # UL1 TR00145.

References

Impact of variation in measurement period for Clinical Quality Measures

Nicholas V. Colin, MA, Raja A. Cholan, BS, David A. Dorr, MD, MS, FACMI
Department of Medical Informatics and Clinical Epidemiology; Oregon Health & Science University, Portland, OR

Introduction: Clinical trials require reliable measures of outcomes; for trials related to quality, electronic clinical quality measures (eCQMs) may facilitate measurement by using evidence-based definitions and data elements found in most Electronic Health Record (EHR) systems. In the context of the national EvidenceNOW initiative, a set of research collaboratives that performed trials focused on heart health-related quality improvement in small-to-medium size primary care practices, four eCQMs were intended to be used as standard measures of practices’ outcomes in quality improvement over time. Since these measures are calculated from EHR data, practices were given standard definitions of eCQMs; however, frequency of measurement varied. In this study, the eCQMs were to be reported on a quarterly basis (rather than yearly – frequently the calendar year - for most standard, federally required reporting initiatives), using a standard one-year measurement period. In Healthy Hearts Northwest – an EvidenceNOW collaborative representing Oregon, Washington, and Idaho – we asked practices to leverage their existing EHRs and Health Information Technology systems to report these data. The ability of EHRs to calculate standard measurement periods varies; many practices reported fixed measurement periods by quarters, by calendar year, or other non-standard durations. The effect of this variation is unknown. In this study, our objective is to understand the impact of varying measurement period on the calculation of outcomes measures for this trial.

Methods: 209 practices were asked to extract and submit four eCQMs from their EHR on a quarterly basis using a standard one-year measurement period. Quarterly submissions were collected via a REDCap survey link. Survey data were processed, and the measurement periods were categorized into non-standard (3, 6, 9 months and other) and standard periods (one-year). We assessed the central tendencies, shape of the distributions, and variabilities across the four measures. Analysis of variance (ANOVA) was conducted to analyze the differences among standard and non-standard measurement period means, and variation among these groups.

Results: Of 209 practices, 191 (91%) submitted at least one measure over three quarters (Quarter 4 of 2015, Quarter 1 of 2016 and Quarter 2 of 2016). Of the 546 total submissions, 173 had non-standard measurement periods. Differences between measures performance with standard versus non-standard measurement periods ranged from -5% to 14.2% between clinics (p<.05 for 3 of 4). ANOVA analysis shows that three of the four measures were statistically different between standard and non-standard measurement periods.

Discussion: We need standardized measures to be implemented for assessing healthcare quality and increasing population health. In practical implementations of eCQMs against EHR data, however, we currently see significant variation due to various sets of transformations, and different organizations of clinical data that currently exist. This includes variation in measurement periods, and therefore we may see differences in eCQM performance. Accounting for these differences based on how practices/clinicians produce eCQM data could be important, for researchers and quality incentive programs. The future collection of EHR data should also include metadata for how it is generated, including details about measurement periods. Ideally, we need standardization in measurement periods in eCQM implementations, across the healthcare system.

Although using native EHR functionality for producing measures did lead to variation in the collected data, the benefit is that we had higher rates of data submission for the official, standardized measures by the Centers for Medicare and Medicaid (Aspirin, Blood Pressure, and Smoking) compared to EvidenceNOW’s novel measure (Cholesterol). Other EvidenceNOW cooperatives extracted data from participating practices’ EHRs and computed the measures in a central quality measure calculation registry. In Healthy Hearts Northwest, we avoided this data extraction process and instead focused on the concept of teaching the practices how to sustain the methodologies they implemented throughout the project.
**Analysis of a Clinical and Translational Science Institute Pilot Grant Competition**

Noshir Contractor, PhD¹, Alina Lungeanu, PhD², Yulia Tyshchuk, PhD³, Griffin Weber, MD, PhD³

¹Northwestern University, Evanston, Illinois; ²The Pennsylvania State University, University Park, Pennsylvania; ³Harvard University, Boston, Massachusetts

**Abstract**

In this poster we describe our approach towards evaluating teams participating in an internal funding competition for $50,000 research pilot grants, conducted within Harvard in 2009 by its Clinical and Translational Science Institute. Of 37,266 eligible faculty, 1,469 formed 458 teams that submitted proposals. Peer-review narrowed this to 99 teams that were invited to in-person interviews, and 65 were awarded funding. We developed a Multi-Theoretical Multilevel (MTML) model to create an integrated explanatory framework to understand collaboration at multiple levels: (1) Individual (actor) level, such as academic rank and gender of each faculty member; (2) Relational (dyad) level, such as prior collaboration between team members; and (3) Higher Order (ecosystem) levels, such as connections between teams. Using the MTML model, we were able to study three distinct phases of the pilot grant process: (1) In the Team Assembly Phase we looked at which investigators chose to collaborate on a proposal and whether new collaborations formed; (2) in the Peer Review Phase, we looked for characteristics of the teams that were awarded funding; and, (3) in the Post-Award Phase, we followed all teams for five years to determine both the impact of funding on the awarded teams and the impact of applying on the non-funded teams. We developed a two-part analytical approach: (1) “Random Teams” compared actual teams that applied for funding to matched virtual teams consisting of randomly selected faculty who did not apply; and (2) “Random Networks” used exponential random graph models (ERGM) to study the applicants and awarded teams in more depth. Our findings include: (1) most teams consisted of a small core of prior collaborators joined by several new faculty, (2) faculty of the same gender were more likely to collaborate on proposals, and (3) non-funded teams continued to collaborate almost as much as funded teams. (This work was supported in part by grants NIH/NIGMS U01GM112623 and NSF/SciSIP #1360042.)

![Figure 1. Comparison of Random Teams and Random Networks.](image)

In both cases, thousands or millions of random controls are generated to develop probability distributions to compare to the actual observed teams and networks. Although Random Teams and Random Networks use a similar approach, they test slightly different hypotheses and return complementary results.

**References**

Improving Identification of Non-Valvular Atrial Fibrillation (NVAF) patient cohorts using Natural Language Processing and SNOMED CT

Peter L. Elkin, MD, MACP, FACMI, FNYAM1, Sarah Mullin, MS2, Gregory Ficheur, PhD, MD3, Daniel R. Schlegel, PhD2, Chris Crowner, MS1, Sylvester Sakilay, MS3, Shyamashree Sinha, MBBS, MBA, MPH1; Gary Brady, PharmD, MBA1, Marcia Wright, PharmD2, Kim Nolen, BS, PharmD1, Sashank Kaushik, MD, MBA1, Jane Zhao, MD1, Buer Song, MD, PhD3, Edwin Anand, MD1

1University at Buffalo, Buffalo, NY; 2Pfizer, New York, NY; 3SUNY Oswego, Oswego, NY

Introduction
Nonvalvular Atrial Fibrillation (NVAF), a type of Atrial Fibrillation (AF) and the most common type of irregular heartbeat, is estimated to affect about 5.8 million people in the United States. Nonvalvular AF results in a five times greater risk of stroke; approximately 15% of strokes annually are due to NVAF. In this retrospective study, the improvement in accuracy of using structured electronic health record (EHR) data (ICD-9) combined with unstructured EHR clinical notes through Natural Language Processing (NLP) techniques (Structured+NLP) was compared with structured (ICD9) data alone, and compared with a clinician review in identifying NVAF.

Methods
The retrospective study cohort included patients between the ages of 18 and 90 with a diagnosis of AF who did not have evidence of significant valvular abnormality. EHR (Allscripts) data from the UBMD faculty practice consisted of 96,681 patients who had both clinical notes and ICD-9 description fields. Following application of the exclusion criteria, electronic models for structured data, using ICD-9 criteria, and for all unstructured data using a high throughput phenotyping NLP system that rapidly assigns ontology terms (including SNOMED CT) to text in patient records, were applied to identify the NVAF population. Two cohorts were created, those identified through structured and NLP and those identified through structured alone. A random sample of 300 patients was selected. In order to create the gold standard, a chart reviewer to review all pertinent records was developed. Each case was reviewed independently by two clinicians and adjudicated by a third clinician when a disagreement occurred. Each NVAF outcome for the respective method was compared to the gold standard by calculating specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV) with their 95% exact binomial confidence intervals. McNemar’s test or the exact binomial test assessed if the two methods performed equally against the gold standard in terms of sensitivity and specificity. Gold standard inter-rater agreement was calculated by Cohen’s kappa with a 95% bootstrapped confidence interval.

Results
The inter-rater agreement for the gold standard was assessed for the 300 cases, each independently reviewed by two clinicians prior to adjudication and moderate agreement was found (0.522 (0.233, 0.748), 0.58 (0.446, 0.706)). Sensitivity, specificity, PPV, NPV, and Cohen’s kappa can be found in Table 1 for each method. A McNemar test comparison of case identification showed that Structured+NLP method’s sensitivity was superior to ICD9 alone (p<0.001). An exact binomial test to assess if the two methods performed equally in terms of specificity against the gold standard failed to reject that the two specificities are different (p=0.317). The positive and negative predictive value was also superior for the Structured + NLP method. Initially, of the 96,681 patients identified in the UBMD database, 2.8% (2722 cases) were identified with NVAF by the Structured+NLP method as opposed to 1.9% for Structured alone (1849 cases) with a difference of 873 cases. This was a significant improvement in identification with Structured+NLP identifying 32.1% more cases than Structured ICD-9 (p<0.001, McNemar Test). Based on the PPV adjusting the true positive rates for both ICD9 and NLP alone this converts to a 36.3% improvement identification of true cases in this NVAF cohort when sample results were extrapolated to the entire data set.

Discussion and Conclusion
The Structured+NLP data extraction method had a higher sensitivity in comparison to Structured data alone, allowing for an increased number of true positive cases to be identified. In addition, Structured+NLP identified 32.1% more NVAF cases than Structured data alone. The enhanced true positive case detection was improved by 36.1%, and the cost of false positive case detection was reduced by 63.7% on average which will reduce the cost of implementation. Improved automated detection through Structured+NLP would accurately identify NVAF patients efficiently compared to structured data alone, but also could facilitate an increased enhanced the ability to recruit to clinical trials.

Table 1. Comparison of outcomes for Structured and Structured plus Unstructured data against the gold standard.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Structured</th>
<th>Structured+NLP</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>.773 (.68, .79)</td>
<td>1 (.979,1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Specificity</td>
<td>.47 (.258, .65)</td>
<td>.444 (.279, .619)</td>
<td>0.317</td>
</tr>
<tr>
<td>PPV</td>
<td>.91 (.87, .95)</td>
<td>.93 (.893, .956)</td>
<td>0.007</td>
</tr>
<tr>
<td>NPV</td>
<td>.215 (.131, .322)</td>
<td>1 (.713, 1)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>kappa</td>
<td>.156 (.041, .271)</td>
<td>.585 (.414, .733)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Semi-supervised Encoding for Outlier Detection in Clinical Observation Data

Hossein Estiri, PhD,1,2,3 Shawn N Murphy, MD, PhD1,2,3
1Harvard Medical School, Boston, MA; 2Massachusetts General Hospital, Boston, MA; 3Partners Healthcare, Boston, MA;

Introduction

It is possible, if not common, to come across biologically implausible values in EHR data (e.g., a negative value for hemoglobin A1c, a BMI of 400, etc.) that are extremely unlikely to signify the “truth” about a patient at a given clinical encounter. Systematic detection of such plausibility issues in EHR data are difficult for two reasons. Even if such global cut-off ranges are available, they may not fully capture biologically implausible values due to an inherent nonlinearity in observation data – i.e., a weight record of 700 lbs. may be plausible for a 50-year-old Male, but certainly not for a 5-month-old. Predictive learning (a.k.a. unsupervised learning) techniques provide low-cost opportunities for constructing adaptive data representations that can be used to evaluate EHR data plausibility.

Methods

Autoencoders are unsupervised Neural Network algorithms that automatically learn features from unlabeled data. One of the applications of autoencoders is in outlier detection by identifying data points that have high reconstruction errors. We hypothesized that training the encoder function on a small random sample of data may result in construction of a more precise “exemplar” representation of the underlying data distribution. We call this implementation semi-supervised encoding, or superencoding. We used data on 28 clinical observations, including records on laboratory results and vital signs from Research Patient Data Registry (RPDR) from Partners HealthCare, for which we obtained global gold/silver-standard cut-off ranges to calculate sensitivity and specificity. We developed and tested 864 (12 × 6 × 4 × 3) superencoder forms per each observation from combinations of 12 sets of input features, six network architectures, 4 sampling ratios, and 3 reconstruction error cut-off values for outlier detection.

Results

We found that almost for all groups of observations values, a superencoder (i.e., the semi-supervised encoder) performed one of the best-in-class outlier detection performances, supporting our first hypothesis. We also found that most of the best-in-class outlier detection performances only included a minimum set of features and had simple network architectures.

Figure 1. Scatter plot of direct bilirubin (DBil) values color-coded by reconstruction error from a superencoder trained on 10% of the data with 2 features (the numeric observation value and age at observation), and a 3-hidden-layer architecture. The Y-axis represents the observation value. Dark red points represent the best input data reconstruction. The dark blue point represents the most extreme outlier.

Conclusion

We have demonstrated the utility of semi-supervised encoders (superencoders) for outlier detection through compressing the data into an exemplar distribution learned from supposedly less noisy random samples. Generally, we have found that simple superencoders can produce acceptable exemplar compressions of clinical observation data and outlier detection performances. However, due to the diversity of distributions in clinical observation, a global solution may not be reachable.

Funding

This work was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-04608) for development of the National Patient-Centered Clinical Research Network, known as PCORnet, and NLM training grant T15LM007092.
Mapping Laboratory Data to LOINC Codes Using Distributions of Values

Laura A. Evans¹, Matvey B. Palchuk, MD, MS¹²
¹TriNetX, Cambridge, MA; ²Harvard Medical School, Boston, MA

Introduction

LOINC is widely considered to be the standard for coding laboratory test results. Unfortunately, typical labs do not include LOINC codes and so individual healthcare organizations (HCOs) need to perform manual mapping to achieve interoperability of lab results - a process which is both time-consuming and tedious¹. A number of approaches have been proposed to make this process easier, including relying on machine learning techniques². However, approaches to date have primarily focused on using the names given to labs to create automated mapping systems. We propose using lab values to provide additional guidance in the mapping process.

Methods

We used data from three HCOs, all large academic medical centers in the US, none of whom use LOINC for their internal lab codes. We chose three labs to investigate: glucose, creatinine, and sodium. We obtained all lab values from each of the HCOs for every lab whose name contained one of our terms (for example, “Glucose Blood” or “Glucose U”). This produced 221 distinct labs for glucose, 63 for sodium, and 167 for creatinine. We dropped any lab with fewer than 1,000 observations and points over the 95th percentile. The average of values for these labs obtained from 32 HCOs on TriNetX network served as our gold standard. From these sets, including the gold standard, we plotted volume of observations over lab values (see Figure 1) and calculated the median, standard deviation, skew, and kurtosis. We then used these four datapoints to characterize the shape of the value distribution. We clustered the plots based on their shapes. The clusters grouped together the labs that have value distributions that look most like each other. We performed the same analysis for the combination of all labs for all 3 terms, to determine if we can distinguish between labs on this more complicated set.

Results

All labs that share a cluster with a gold standard are tentatively mapped to the LOINC code associated with the gold standard. If no gold standard code is in a cluster, the lab in that cluster remain unmapped. Results are subsequently reviewed by hand. Having scored each predicted mapping as correct or incorrect, we evaluated this method by calculating such accuracy metrics as recall, precision, and F-score. The algorithm demonstrated strong performance when comparing labs for a single term to each other (e.g., glucose in blood F-score 0.7), but weaker performance when comparing labs for multiple terms.

Conclusion

We described an approach to use the shape of the value distribution to aid in mapping labs to LOINC. We realize that shape alone will not achieve a perfect mapping, but are considering it to be a useful addition to the process. Further work is needed to extend this approach to be functional for multi-term comparisons. However, for single-term comparisons, our algorithm can provide significant assistance to the manual mapping process. Additionally, the success of the algorithm described in this work suggests that automated methods for lab mappings are possible.

References

Identification of Risk Factors for High Utilization of Healthcare in Diabetic Patients Using an Integrated Medical Records Database.

Anika Ghosh, BS, Kathryn Jackson, MS, Kate Balsley, MPH, Abel Kho, MD, Marc Rosenman, MD, Theresa Walunas, PhD; Northwestern University, Center of Health Information Partnerships, Chicago, IL

Introduction: Healthcare costs are extremely high in the U.S., and only half of the population accrues 97% of the costs. The Transforming Clinical Practice Initiative (TCPI) funded by the Centers for Medicare and Medicaid Services aims to improve care and reduce costs by providing health care providers resources to help improve quality of care. A key health care focus for improvement is diabetes. Approximately 9.4% of the US have diabetes, and many present with multiple comorbidities that can require significant management and lead to poor disease control and high utilization of emergency department (ED) visits. To improve care for patients with diabetes in Chicago, we plan to develop a model that predicts attributes correlated with high utilization in diabetic patients, which may help us identify strategies to prevent or mitigate high utilization and thereby reduce costs.

Methods: This study used electronic health record (EHR) data from the Chicago HealthLNK Data Repository (HDR), which is a city-wide database of de-duplicated, merged patient records of demographics, ICD9 codes, labs, and medications from seven institutions in the Chicago area from 2006-2012. We defined high utilization as having four or more ED visits in twelve months. We used the 250.x ICD9 codes to define diabetes. Fragmentation of care was defined as a visit at more than one HDR institution. ICD9 codes were categorized into phenotypes (using the Phenome Wide Association Studies [PheWAS] catalog) to identify comorbidities and outcomes. To ensure that we had the first instance of high utilization, we made sure we had a year of data after identification of a diabetes diagnosis, but before the first ED visit that led to high utilization in order to assess factors that could impact high utilization. We used logistic regression in SAS 9.4 to identify predictors of high utilization and used chi square analysis and t-tests to determine significant variables (*) at a p<.05 level.

Results: We had 87,569 patients in our study cohort that met our definition of diabetes, and 4,984 were identified as high utilizers. Preliminary results show high utilizers with diabetes are more likely to have depression, substance addiction problems, and pain (joint or nonspecific chest) compared to a diabetic non high utilizers (Table 2). They are also more likely to have worse outcomes from diabetes such as diabetic retinopathy and diabetic ketoacidosis (Table 2). High utilizers are on average younger than non-high utilizers, have a higher occurrence of fragmented care and are also more likely to be African American and the beneficiaries of public insurance (Table 1).

Conclusions: Diabetic patients who are high utilizers of health care are more likely to have complications from diabetes and have comorbidities like pain, addiction, and depression that make it hard to control their diabetes and lead to high utilization. Future work will focus on developing a model to identify attributes that predict high utilization, including comorbidities, disease severity, and healthcare usage as well as social and behavioral determinants of health. These predictors will be used to inform clinical interventions and decision support and help healthcare providers develop strategies for population health management that can improve care and health outcomes as well as reduce costs for patients with diabetes.

Table 1: Patient Characteristics

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>High Utilizers</th>
<th>Non high utilizers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Age</td>
<td>56.55 years</td>
<td>62.2 years</td>
</tr>
<tr>
<td>African American</td>
<td>72.50%</td>
<td>39.00%</td>
</tr>
<tr>
<td>Fragmented Care</td>
<td>46.00%</td>
<td>27.00%</td>
</tr>
<tr>
<td>Medicare/Medicaid</td>
<td>75.40%</td>
<td>58.00%</td>
</tr>
</tbody>
</table>

Table 2: Outcomes and Comorbidities

<table>
<thead>
<tr>
<th>Conditions</th>
<th>High Utilizers</th>
<th>Non high utilizers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonspecific chest pain</td>
<td>29.99%</td>
<td>10.44%</td>
</tr>
<tr>
<td>Pain in Joint</td>
<td>26.40%</td>
<td>13.54%</td>
</tr>
<tr>
<td>Depression</td>
<td>12.70%</td>
<td>4.60%</td>
</tr>
<tr>
<td>Substance Addiction</td>
<td>10.05%</td>
<td>1.50%</td>
</tr>
<tr>
<td>Diabetic Retinopathy</td>
<td>6.11%</td>
<td>3.97%</td>
</tr>
<tr>
<td>Diabetes Ketoacidosis</td>
<td>1.79%</td>
<td>.51%</td>
</tr>
</tbody>
</table>
Review of Health and Fitness Apps for Hands-Free Voice-Activated Assistants

Ashley C. Griffin, MSPH¹; Dasha Selzneva¹; David Gotz, PhD¹; Arlene E. Chung, MD, MHA, MMCi¹
¹University of North Carolina at Chapel Hill Carolina Health Informatics Program, Chapel Hill, NC, USA

Introduction

Hands-free voice-activated assistant (VAAs) and their associated devices have recently gained popularity with the release of commercial products, including Amazon Alexa and Google Assistant with others pending release in 2017 from Apple and Microsoft. VAAs have many potential use cases in healthcare including health education, health tracking and monitoring, and assistance with locating health providers. Since 2014, there has been a rise in applications (apps) for these devices but there are no prior studies that describe the content and characteristics of health and fitness apps for hands-free VAA devices. Thus, we report the first review of health and fitness apps for hands-free VAA devices.

Methods

We searched for commercially available, hands-free VAAs that had app stores. Only two VAAs were available, which were the Amazon Alexa and Google Assistant VAAs. Inclusion criteria included that VAA apps were designated by vendors in the “health and fitness” categories in VAA app stores. Thus, the Alexa Skills Store and Google Assistant website/apps were searched to find VAA apps designated by vendors as “health and fitness” apps. Information was extracted about each app including description and name, vendor, ratings, cost, and developer and security policies. Using a codebook, two reviewers independently coded each app using the vendor’s descriptions and app name into one or more health/fitness categories, intended age, and target audience groups. Apps could be coded into multiple categories. A third reviewer adjudicated any coding disagreements until consensus was reached. Descriptive statistics were used to summarize app characteristics and inter-rater agreement was calculated.

Results

There were a total of 309 apps that met inclusion criteria. 300 apps were identified from the Amazon Skills Store website, and 9 apps were identified from the Google Assistant app. Apps had a release date between 11/6/2015 and 4/19/2017 (date of extraction). Health education apps (87) were the most common type, followed by fitness/training (72), nutrition (33), brain training/games (31), and health monitoring (25). On average, users rated health and fitness apps 2.97 out of 5 stars. All apps were free to enable, though some required an associated account, which may require a subscription fee. The majority of apps (87%) were rated by the vendor as having “mature or guidance suggested” content and were for adults only. 18% of apps could be paired with a mobile app, website, or a device. 31% of apps had a privacy policy and 22% had a developer policy for terms of use. Inter-rater agreement was 91%.

Discussion

Although the VAA apps reviewed in this study were released into the health and fitness category, many did not seem to have a clear health or fitness focus. VAA apps had a wide range of content areas and were mostly targeted towards adults and general audiences. This may be due, in part, to Amazon and Google’s policies that place restrictions on health-related apps such that it limits healthcare delivery through VAAs. VAAs, which are continuously “listening” in the background, raise a number of privacy and security concerns, which consumers may not be aware of or be adequately informed about.

Conclusions

The emerging market of health and fitness apps for VAAs is nascent and does not have many health-focused apps available in comparison to the mobile health app market, where chronic disease, monitoring, and self-management mobile apps have proliferated. Privacy, security, and HIPAA compliance need to be addressed along with the stringent requirements from publishers to allow for the use of VAAs to enable and deliver healthcare securely while maintaining privacy. Further work is necessary to evaluate the usefulness, usability, content, quality, and privacy implications of VAAs and associated apps.
A survey of strategies used by drug information experts to search for evidence about potential drug-drug interactions

Amy J. Grizzle, PharmD1, John Horn, PharmD2, Carol Collins, MD2, Daniel C. Malone, PhD1, Britney Stottlemeyer3, Richard D. Boyce, PhD3

1University of Arizona, Tucson, AZ; 2University of Washington, Seattle, WA; 3University of Pittsburgh, PA

Introduction
Ensuring that medication therapy occurs safely and to the maximum benefit of any given patient is of great interest to clinicians. However, when it comes to potential drug-drug interactions (PDDIs), clinicians often face barriers such as incomplete personal knowledge and alerts with poor specificity. Moreover, inconsistencies in PDDI content across data sources are well-documented. A study of three commercial knowledgebases found that of 8.6 million PDDIs, only 5% were included in all three sources. In prior work we reported that the workflow of individuals who maintain PDDI information generally involves topic identification, evidence search, synthesis, and recommendation. The lack of validated standardized search strategies for PDDI evidence may be a major factor contributing to variability across sources of drug information. We surveyed drug information experts to better understand the various ways that they conduct PDDI evidence searches.

Methods
A convenience sample of 70 drug information experts, including compendia editors, knowledgebase vendors, and clinicians, were invited to complete a survey about how they search for evidence of PDDIs. Those who agreed to participate were sent a link to the 13-item survey. Questions focused on how they develop and conduct literature searches, strategies they utilize, as well as the sources typically used. Surveys were anonymous and participants were compensated up to $20 for their time.

Results
Twenty of the 70 invitees (response rate 29%) completed the survey. The majority (85%) of respondents were drug information specialists, drug interaction researchers, compendia editors, or clinical pharmacists, with 60% (12/19) having over 10 years of experience. Most experts develop (18/20; 90%) and conduct (19/20; 95%) their own search strategies without assistance from a librarian. All 20 participants use PubMed and 11 (55%) include Google Scholar when searching for published scientific articles. Embase, with a more international focus, is used by very few respondents (3/20; 15%). To identify PDDIs, respondents most commonly use the key terms “generic name,” “enzyme name,” “transporter name,” and “drug interaction.” Additional search strategies include features like “find similar” or “related articles” within search engines (15/20; 75%) and implementing backwards reference searches (16/20; 80%). The most common databases used by participants included subscription databases: Lexicomp (9/20; 45%) and Micromedex (8/20; 40%); open access databases: Drugs@FDA (17/20; 85%) and DailyMed (13/20; 65%). Experts reported using a variety of compendia including Facts and Comparisons (8/20; 40%), Top 100 Drug Interactions (7/20; 35%), and Drug Interactions: Analysis and Management (6/20; 30%). The most commonly reported web-based resources included product labels (18/20; 90%), Medwatch, and DailyMed (both 12/20; 60%).

Conclusion
Drug information experts use a variety of keyword strategies and evidence sources while searching for PDDI information. Findings from the survey indicate that experts develop and conduct searches without assistance from a librarian, tend to use PubMed and Google tools, but not Embase, for scientific article searches, and use various keyword strategies and third party references depending on the kind of PDDI evidence they are searching for. Many participants shared specific heuristics and keyword strategies. These data as a whole will be used by us to design candidate standard PDDI search strategies that will undergo rigorous validation.

References
Utilizing medical informatics to expand the reach of comparative effectiveness research

Cynthia Hau¹, MPH, Nilla S. Majahalme¹, BSE, Sarah M Leatherman¹, PhD, Oleg V. Soloviev¹, MS, Patricia A. Woods¹, RN, MSN, Amanda S. Guski¹, MS, Maura A. Flynn¹, RN, MSN, Christal T. Sadatis¹, AS, Jade A. Fiotto¹, MPH, Alison L. Majkut¹, MPH, and Ryan E. Ferguson¹, ScD, MPH

¹Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC), VA Cooperative Studies Program, VA Boston Healthcare System, Boston, MA, USA

Description of the problem: There is a growing interest to leverage healthcare information technology for the conduct of randomized comparative effectiveness research (CER). While real-time analytics using electronic medical record (EMR) data is an essential component and widely practiced, few studies have utilized available EMR systems to conduct CER procedures that are traditionally managed at in-person visits, limiting its reach to patients in remote locations. In an attempt to reach patients across the nation, we extended the application of medical informatics to streamline the entire patient enrollment process.

Specific purposes of the system or project: An informatics-enabled system was implemented. In addition to the typical scope of EMR data collection and analysis, this system is positioned to guide research team with subsequent enrollment steps, allowing us to pre-screen, consent and randomize remote patients from a central coordinating center. This system is currently implemented to a 4-year CER and scheduled to launch in 50 VA medical centers, enrolling 13,500 older veterans nationwide.

Methodology: Our system is structured to analyze real-time patient data from the VA Corporate Data Warehouse (CDW), which comprises 8 main data domains including contact information, background characteristics, healthcare utilization, vital signs, anthropometrics measures, laboratory results, pharmacy records, and health status (indicated by inpatient, outpatient, and vital records). These data are extracted as a study site is launched and transformed to aggregate results using pre-defined statistical algorithms. As shown in the figure below, the study-relevant EMR results are loaded to various medical data communication platforms (e.g. Computerized Patient Record System, Computerized Provider Order Entry system, and Mail and Call Center Information Management Systems) in order to guide the specific research study group to the appropriate next steps.

Preliminary results, Discussion, & Potential implications: Using medical informatics to operationalize the entire patient enrollment workflow has greatly expanded our capacity to include nationwide patients in our CER. During the first 9-months of active enrollment, we successfully launched this innovative workflow in 3 VA medical centers, enabling us to identify 2,712 eligible patients, and randomize 330. Consequently, we expect exponential growth in the enrollment rate when this workflow is launched at other targeted VA medical centers in the coming months. We believe this abstract illustrates a meaningful application of medical informatics to operationalize the entire patient enrollment process. The concept of embedding patient enrollment process to current EMR infrastructure, such as our EMR-integrated enrollment system, can potentially be applied to other medical centers to expand their enrollment capacity to include remote patients in CERs.
Evaluation of Abbyy FineReader Optical Character Recognition Tool Applied to Scanned Outside Medical Records to Facilitate Clinical Research

Julie Hom BS, Janet Nikowitz AA, Rebecca Ottesen MS, Russ Hughes BS, Joyce Niland PhD
City of Hope, Duarte CA

Introduction  Patients’ past medical histories often include numerous scanned records from outside hospitals within the Electronic Medical Record (EMR), particularly at tertiary centers such as City of Hope (COH). As these scanned documents are rendered as images, it becomes an arduous task to locate and codify key data elements required for clinical research.[1] Further, the images are not standardized nor structured, and therefore are not amenable to applying data mining to facilitate data abstraction. Reviewing hundreds to thousands of pages as images is an extremely impractical, time consuming process to capture research data. We evaluated the Abbyy FineReader (AFR) Optical Character Recognition (OCR) software, an application that can transform scanned images into searchable documents[2], to support text mining to indicate the presence of research interest content within a corpus of scanned documents. Tafti et al. compared 4 OCR systems to process various types of documents, and AFR yielded promising results.[3] We applied AFR to transform scanned image documents into minable text, to attempt to reduce abstraction time. We chose a cohort and corpus from a prior project where manual abstraction had already been performed and validated, to serve as the ‘gold standard’ results. If successful this approach could greatly decrease the time and effort to read through and process scanned outside medical records for clinical research.

Methods  Our cohort consisted of 189 subjects diagnosed with diffuse Large B-Cell Lymphoma, for whom performance status (PS) at diagnosis was a required data element to analyze prognostic factors for recurrence/survival. Because PS must be captured very close to diagnosis, this information often is located in outside scanned medical records. Leveraging our COH data warehouse infrastructure and using the Informatica Powercenter Extract Transform Load (ETL) tool, a queryable table was created in our data mart containing all documents for the selected patients (see Figure 1 for process flow). We refined our search to a subset of 4 relevant record types to yield a tenable corpus of ~9,600 records (see Table 1). OCR was applied to TIFF files to transform them into FineReader documents, which were then saved as PDFs to enable text searching and bookmarking of identified data terms. Text mining was applied to the PDF files to ‘tag’ pertinent key words indicative of the presence of PS information, including “performance status”, “Karnofsky”, “KPS”, and “ECOG”.

Results  Transforming scanned images into FineReader PDF documents did indeed allow the application of text mining to tag keywords of interest to facilitate data abstraction, as a ‘pre-filter’ to identify those pages that warranted review for coding of PSs. Table 1 shows that Outside Progress Notes yielded the majority of “hits” for PS, while the other documents types yielded minimal results of interest. Evaluating AFR results against the manually curated gold standard showed only a very small fraction of missed PS (only 19 instances in 4,545 pages reviewed for the first 25 patients); while the Subject Matter Expert was unable to locate 32 of the AFR ‘hits’ after thorough manual review. A complete analysis of precision and recall is underway, examples show that the information retrieval time can be decreased from several hours per patient when performing purely manual review, to just a few minutes time with access to the AFR PDF documents.

Discussion  Using AFR to tag documents for presence of relevant research information via keywords within the scanned images allows for a highly focused data review, freeing time for other data elements requiring more human interpretation. Reducing the corpus of records requiring manual review has excellent potential, allowing the abstractor to focus only on those pages with positive ‘hits’. Refining the key words used could further improve results, and more complex clinical research concepts would certainly involve a larger set of terms. Naturally, poor quality scans or embedded hand-written notes, decreases the performance for such technology.

Further refinements to reduce manual effort could include interpreting dates of capture to select only those closest to diagnosis, and up until the first treatment. Integrating Natural Language Processing (NLP) software to identify contextual clinical concepts within the scanned medical records is being evaluated as well. It is clear that this methodology can potentially yield a huge reduction in information review requirements, by up to ~98%.

References
Machine learning on EHR data for prediction of emergency visits and hospitalization during cancer radiotherapy
Julian C. Hong, MD, MS, Manisha Palta, MD, Jessica D. Tenenbaum, PhD
1Duke University, Durham, NC, USA

Introduction
Patients with cancer undergoing radiation therapy (RT) may require emergency department (ED) evaluation or hospitalization, impacting treatment outcomes and increasing healthcare costs. Early identification may direct supportive care to prevent these events. The objective of this study is to apply machine learning (ML) to pre-treatment and treatment EHR data to predict ED visits and hospitalization during treatment.

Methods
We identified 8,462 outpatient RT courses for adult patients at Duke from 2013-2016. Structured EHR data including demographics, encounters and vitals in the past year, labs in the past four weeks, medical history by ICD, and medications at the start of RT were extracted from Duke’s enterprise data warehouse. RxMix was used to unify medication RxNorm names and MeSHPA classes. Treatment information included number of RT treatments, RT technique, and concurrent or recent systemic therapy. Random training (75%) and test (25%) sets were generated. Random forest (RF), gradient tree boosting (GTB), and support vector machine (SVM) were trained on the training set for ED visits or admissions and validated on the test set. The ML approach with greatest AUC was also used to compute a model with only disease and treatment-related variables. GTB hyperparameters were tuned by 5-fold cross-validation and SVM inputs were pre-processed by centering, scaling, and removal of low variance variables.

Results
GTB (AUC 0.799), RF (0.767), and SVM (0.779) were predictive for ED visits or admissions and had greater AUC than GTB trained on only disease and treatment-related factors (Figure 1). GTB variable importance identified treatment factors (number of RT treatments, concurrent systemic therapy) and pre-treatment encounters (number of and recentness of ED visits, admissions), vitals (weight loss and pain score), and labs (albumin).

Conclusions and Future Directions
ML using structured data at the start of treatment can predict patients requiring ED evaluation or admission during cancer RT and may assist direction of supportive care. Work is ongoing to apply natural language processing to extract biomedical concepts from clinical documentation. A prospective trial investigating ML-assisted direction of increased clinical assessments during treatment is planned.

Figure 1. Receiver operating characteristic curves for ML models.

References
Selecting EHR-driven Recruitment Strategies: An Evidence-Based Approach

Daniel R. Hood, MPH¹, William K. Barnett, PhD¹,², Peter J. Embi, MD, MS¹,²
¹Regenstrief Institute, Inc.; ²Indiana University School of Medicine. Indianapolis, IN

Abstract

Participant recruitment for research is a persistent bottleneck that can be improved by leveraging EHRs. We assessed the literature and engaged domain experts in an iterative manner to inform a decision-making approach to selection of EHR-driven strategies. We present requirements, a proposed decision matrix, and future directions.

Introduction

Low rates of participant recruitment for clinical research are a major barrier to advancing science and discovery (1). The Trial Innovation Network (trialinnovationnetwork.org), specifically the Recruitment Innovation Center (RIC), funded by NCATS, are developing methods to improve recruitment. EHR-based recruitment approaches are increasingly being used, however, little guidance exists to guide use of one strategy over another. We developed an evidence-based approach to the selection and use of EHRs for patient recruitment.

Methods

Based upon a review of the literature related to EHR-driven recruitment and collection of firsthand feedback using some of these methodologies from our team’s content experts, we developed and categorized a list of EHR-driven recruitment approaches. We documented literature-derived considerations for the use of certain approaches (2-5). We then conducted a series of group discussions with content experts to elicit their experiences and internal heuristics for the selection of specific EHR-driven recruitment approaches. Field notes were taken and used to create a list of key questions to inform the use of one approach over others. From this, we developed a decision matrix and presented back to the content experts for member-checking and refinement of the decision approach.

Results

EHR-based recruitment strategies were categorized into six groups: (a) patient/PHR-directed alerts; (b) clinician-directed point-of-care alerts; (c) researcher-directed alerts; (d) registry-driven recruitment; (e) general EHR cohort identification for asynchronous participant contact; (f) patient-directed kiosk-based alerts. Key questions to guide decisions fell into four categories: (a) study population characteristics; (b) study protocol characteristics; (c) clinical setting considerations; (d) EHR-specific considerations. Based upon these, we developed an initial decision matrix.

Conclusions

EHR-driven approaches to participant recruitment have potential to improve recruitment rates. This is the first attempt we are aware of to create a guide to the selection and use of EHR-based recruitment approaches based upon best evidence (i.e. literature and expert opinion). More research is needed to to validate this approach as applied to real-world studies. We are actively evaluating this approach via studies utilizing CTSA RIC services.

Acknowledgements: Thanks to our colleagues in the NCATS supported RIC for their input and contributions to this effort. Support provided by: NIH/NCATS: 1U24TR001579-01

References

Identification, Remediation and Knowledge Management of Data Quality
Issues in PCORnet: Experiences in the Obesity Demonstration Projects

Casie E. Horgan¹, MPH, Jessica L. Sturtevant¹, MS, Jane Anau², Juliane S. Reynolds, MPH¹, David Arterburn², MD, MPH, Jason P. Block, MD, MPH¹, Jeffrey S. Brown, PhD¹
¹Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; ²Kaiser Permanente Washington Health Research Institute, Seattle, WA

The National Patient-Centered Clinical Research Network (PCORnet) is a distributed research network (DRN) which facilitates rapid, large-scale clinical trials and observational research. The PCORnet DRN currently includes clinical data from over 128 million patients¹. PCORnet uses a common data model (CDM) to organize and standardize data across Network Partners (NPs) with heterogenous source data systems. The PCORnet CDM consists of 15 domains and over 100 variables.

The PCORnet Obesity Demonstration Projects were the first observational research studies to use the PCORnet CDM. While NP data is assessed in an overall data curation process by the PCORnet Coordinating Center, a more specified assessment of study-specific variables and cohorts is required to ensure data consistency and usability amongst study sites. A reusable, modularized program was developed to rapidly generate counts, frequencies and distributions of important study variables in a process called Study Specific Data Characterization (SSDC). Through SSDC, issues were identified via communications with NPs, execution of distributed SAS programs, evaluation of program logs, and detailed examination of returned data. Data and related policy issues were tracked and later analyzed. Six recurring themes and methods for remediation (Table 1) were identified and shared with PCORnet stakeholders.

Table 1. Recurring issues identified through PCORnet obesity demonstration projects

<table>
<thead>
<tr>
<th>Issue Theme</th>
<th>Resolved Issues</th>
<th>Unresolved Issues</th>
<th>Remediation Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prescribing table implementation and data quality issues (e.g. lack of granularity in RxNorm coding prevented differentiation between drugs)</td>
<td>44</td>
<td>11</td>
<td>Require study-specific characterization of prescribing table prior to using data for research.</td>
</tr>
<tr>
<td>SAS execution issues (e.g. UNIX issues with program case-sensitivity prevented successful execution)</td>
<td>25</td>
<td>1</td>
<td>Use standard program format and programming rules when developing distributed programs so NPs are familiar with execution.</td>
</tr>
<tr>
<td>Other CDM table implementation and data quality issues (e.g. incorrect unit conversion for vital measures)</td>
<td>13</td>
<td>7</td>
<td>Require study specific characterization of data. Don’t assume previously approved data will have been analyzed for study-specific variables and code lists. Maintenance of PCORnet data issue repository will allow for easier error identification and remediation.</td>
</tr>
<tr>
<td>Lab table implementation and data quality issues (e.g. truncation of lab results indicated lab values of ’0’ or ’1’ only)</td>
<td>7</td>
<td>6</td>
<td>Require study-specific characterization of lab table prior to using data for research.</td>
</tr>
<tr>
<td>Delayed/incomplete response to SSDC queries due to governance or failure to pass foundational PCORnet data curation</td>
<td>8</td>
<td>5</td>
<td>Build time into timelines to allow for study specific data characterization, remediation, and subsequent reassessment. NPs must have policy on data review/return</td>
</tr>
</tbody>
</table>

Continued tracking and analysis of data issues across network projects will allow for more coordinated knowledge management of lessons learned and will improve network processes and data quality. The issue themes and resolutions in the Demonstration Projects will be shared with future investigators to allow for increased study efficiency and understanding of data usability and consistency. In highlighting recurring problems, the Network can focus resources on areas requiring improvement and inform eligibility assessments and programming rules for future queries.

References

Developing Efficient, Reusable SAS Programs in PCORnet: PCORnet Modular Program 1 (PMP1)

Jensen N. Hu, MPH\textsuperscript{1}, Julia A. Fearrington, MPH\textsuperscript{1}, Jessica L. Sturtevant, MS\textsuperscript{1}, Casie E. Horgan, MPH\textsuperscript{1}, Érick Moyneur, MA\textsuperscript{2}, Dave Martin, MA\textsuperscript{2}, Darcy M. Louzao, PhD\textsuperscript{3}, Jeffrey S. Brown, PhD\textsuperscript{1}

\textsuperscript{1}Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; \textsuperscript{2}Statlog Consulting Inc., Ottawa, ON, Canada; \textsuperscript{3}Duke Clinical Research Institute, Durham, NC;

Introduction:

The National Patient-Centered Clinical Research Network (PCORnet) is a Distributed Research Network (DRN) that relies on a Common Data Model (CDM) to standardize data and enable distributed querying in support of comparative effectiveness research, pragmatic clinical trials, and observational research. The PCORnet Coordinating Center (CC) creates SAS programs designed to execute against CDM data from 81 Network Partners (NPs) which are organized into 13 Clinical Data Research Networks (CDRNs). Development of SAS programs for use across a large, technically-diverse network consisting of different SAS versions (9.3 & 9.4), operating systems (LINUX/UNIX, Windows), and data storage platforms, requires a focus on reusability, efficiency, and reliability.

The first reusable, analytic SAS program, PCORnet Modular Program 1 (PMP1), identifies patient cohorts defined by diagnosis and/or procedure data. PMP1 was developed by leveraging a Food and Drug Administration (FDA) Sentinel System routine querying tool. Development of PMP1 version 2 was based on input from a NP working group and focused on creation of a report to improve output interpretability with additional aims of improving efficiency and usability. As of December 31, 2017, PMP1 has supported 13 research requests.

Methods:

Gathering specifications for PMP1v2 involved a working group of 3 CDRNs—PaTH, Mid-South, and REACHnet—consisting of network principal investigators, project managers, data analysts, and response reviewers. The working group provided feedback on content and format of the report to allow for easier review by constituents. Building of SAS infrastructure involved adding and modifying SAS code to create the report and improve efficiency. Testing was performed using a PMP1v1-documented and tested diagnosis-based phenotype query on 13 NPs selected to represent PCORnet computing and infrastructure heterogeneity. A survey was distributed after release of PMP1v2 to gauge network agreement regarding interpretability of output, facilitation of internal review and governance process, and usability of PMP1.

Results:

Eleven of the 13 NPs returned output within the standard 2-week response period. Responding sites represented UNIX/LINUX and Windows operating systems, different SAS versions, and variable patient population sizes. On average, the execution time between PMP1 v1 and v2 decreased by 50%. Of the 22 individuals, representing 8 CDRNs, that responded to the survey, 60% agreed that PMP1 usability improved, 54% agreed that enhancements helped governance process and expedited return of output, and 86% agreed that enhancements improved the interpretability and review of output. PMPv2 was released on June 12, 2017 and as of December 31, 2017 has supported 3 requests.

Discussion:

Through collaboration with NPs, updates to reusable SAS programs can benefit the DRN. PMP1v2 improved interpretability of results by creating a summary report and improved efficiency by dramatically improving program performance runtime. Dedicated resources and understanding of competing priorities of internal and external partners are necessary when developing SAS programs in a DRN. Future enhancements and testing will remain necessary as Network preferences and governance evolves.
Heterogeneity of Height & Weight Source Values in Clinical Data

Nnenna Ibeanusi, MSPH/MSIS¹, John P. Borsi, BS¹, Wei Yao MS¹
¹IBM Watson Health, Cleveland, OH

Introduction

An invariable feature of Electronic Health Record (EHR) data is that it is messy, and for height and weight data, the building blocks of body mass index (BMI) calculation, this holds true.¹,² There is a lack of standard observation value formats and units of measurement ( uom ) for height and weight values in EHR data. Not only are these values heterogeneous, but they are also hard to interpret and sometimes conflicting. The consequences of this heterogeneity of observation value formats and uom include lack of interoperability of clinical data across EHR's. But more importantly, for public health research, this often leads to data that is thrown away, due to misinterpretation, or that the data that is used is not interpreted correctly and results in poor BMI measurement values. Although a feature of EHR data, heterogeneity of source data should not be an impediment to using BMI data. Rather, being able to account for and work with heterogeneous formats are necessary to improve the volume, quality and veracity of EHR data.

Data Source & Methodology

Height and weight values and uom were collected from the IBM Explorys Therapeutic data set. This platform holds integrated EHR data across multiple health systems and across multiple source systems for more than 54 million patients.³ Records with missing observation values were removed from the analysis. There were over 1,000 unique observation value formats and over 100 unique uom found across the data set. Given the fact that the platform is real-time such that new data and EHR sources are continually being added, further increasing the diversity of uom values, a deterministic method of text manipulation was applied to the grouping logic that defined a standardized uom field. Distributions for height and weight were aggregated across all organizations and for each organization to understand the range and frequency of observation values for each standardized uom before and after the Height/Weight cleaner (HWC) algorithm, which interprets the heterogeneous observation value formats and determines a correct observation value and uom, is implemented.

We propose an approach using the observation value format and the standardized uom to interpret the true H/W value. We determine the conditional probability of a given unit interpretation given the raw input data and patient demographics. Prior distributions are derived using a Gaussian Mixture Model. The posterior probability estimates are used to determine the appropriate unit interpretation. This approach handles common transformations as well as measurement errors and typos.

Conclusion

Heterogeneity of data in EHR systems is characteristic of the data. Efforts to clean and standardize data to determine appropriate values are necessary, particularly with data elements such as BMI that drive value to understand patient populations and help derive insights for health outcomes.

References

Use of Clinical Quality Measures in a Multi-Site Integrated Electronic Health Record Database as Measurements of Population Health

Kathryn L Jackson, Kate Balsley, MPH, Marc B Rosenman, MD, Abel N Kho, MD, Theresa L Walunas, PhD
Northwestern University, Center for Health Information Partnerships, Chicago, Illinois

Introduction
The Centers for Medicare and Medicaid Services (CMS) Quality Payment Program (QPP) requires physicians to report on clinical quality measures (CQM) based on data from Electronic Health Records (EHR). These individual-level quality metrics also align with important public health concerns and initiatives, and could provide a means to measure population health. However, intricacies of health system EHRs and data warehouses may require use of clinical notes or other non-structured data elements in order to generate these metrics accurately. Additionally, the data necessary to accurately measure care quality may be fragmented across multiple institutions in which a patient receives care. Due to HIPAA regulations around protected health information (PHI), sharing data containing identifying information (e.g., clinical notes, free-text data elements, full visit dates) is problematic, hindering our ability to create and utilize integrated data sources to measure the health of larger populations. If structured data containing no PHI could be used in generating these quality measurements, we could better assess quality of care across institutions and, therefore, across populations when multisite data are integrated. Using this framework, we can then attempt to measure health of a city using an integrated data set.

Methods
The HealthLNK Data Repository (HDR) is an integrated city-wide database containing patient-level EHR data from seven institutions in Chicago. The HDR uses a HIPAA compliant hashing algorithm to create a unique ID for each patient, allowing for the creation of a merged and de-duplicated patient data set, without the need to share PHI. Structured data elements (demographics, ICD-9/CPT codes, medication orders, and laboratory results) were collected using a common data model. For this project, we use the HDR to assess the feasibility of calculating population health measurements in an integrated database using CQMs and explore what additional data are necessary to collect in order to enhance these measurements.

CQMs chosen by CMS for the QPP for internal/family medicine in the National Quality Strategy and Community/Population Health and Effective Clinical Care domains were selected for evaluation (https://qpp.cms.gov). Data elements required for reporting were recorded for each numerator and denominator. Numerators and denominators were evaluated for feasibility of use (yes/no) based on the data elements available in the HDR (e.g. diagnosis and procedure codes, medication orders, laboratory results). For non-feasible numerators and denominators, non-available data elements were recorded and categorized based on probable location of data in the EHR (e.g. clinical notes, behavioral notes, behavioral screening, intervention planning). Results were examined for the entire set and stratified by measure type (process or outcome). CQMs were calculated for feasible measures and compared against national benchmarks.

Results
A total of 34 CQMs (n=5 (15%) outcome; n=29 (85%) process) were evaluated. Overall, 13 (38%) CQMs met feasibility criteria for both numerator and denominator; 3 (of 5; 60%) outcome and 10 (of 29; 34%) process measurements met these criteria and therefore could be used as city-wide measurements of population health. An additional 13 process measures met the feasibility requirements for the denominator alone. The final 8 CQMs could not be calculated, as both numerator and denominator were infeasible. All 34 CQMs included a corresponding CPT-II or HCPCS code for inclusion in the numerator, however, these codes are often not collected in the EHR. If these codes were consistently collected, each numerator could be calculated. In absence of CPT-II/HCPCS codes, information needed for those measures not meeting feasibility requirements likely would be found in behavioral screening results (n=7), intervention plan records (n=6), and/or clinical/exam notes (n=3). Overall, calculated CQMs for feasible measures were within range of national benchmarks (e.g. 54.8% for poor diabetes control and 55.9% for high blood pressure control).

Discussion
Structured and semi-structured data elements found in EHRs can be used to develop measurements of overall population health; however, inclusion of patient-level CPT-II/HCPCS codes, behavioral screening, and intervention documentation as structured fields would greatly enhance our ability to measure population health based on CQMs. Additionally, calculated CQMs were within range of national benchmarks, suggesting that population health can be accurately measured using the quality framework in integrated data sets. Currently, a refresh of the HDR is underway in which additional procedure data, including CPT/HCPCS data, are being collected from all seven institutions; the refresh will ensure more accurate results and an even better snapshot of health care quality at a population level. Future research will explore potential methods of collecting non-structured data into structured fields which can be added to the current common data model.
A Systematic Analysis of the Comorbidity Patterns of Schizophrenia using a Nationwide Health Insurance Claims Database

Di Jin, MS¹, Nathan Palmer, PhD², Isaac S. Kohane, MD, PhD², Peter Szolovits, PhD¹, Kun-Hsing Yu, MD, PhD²

¹Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA; ²Department of Biomedical Informatics, Harvard Medical School, Boston, MA

Introduction
Schizophrenia affects more than 3.2 million patients in the US¹, costing approximately $155.7 billion per year². Many comorbidities have been reported, including anxiety disorders, depression, and bipolar³. However, due to the complexity of the disease manifestations, a comprehensive analysis on the comorbidities of schizophrenia is lacking. To better characterize the clinical phenotypes of schizophrenic patients, we conducted a systematic analysis of the comorbidity patterns of schizophrenia using a nationwide health insurance dataset, delineated the phenotypic differences across the common subtypes, and characterized the risk factors of this disabling disease.

Materials and Methods
Using un-identifiable claims data from a nationwide U.S. health insurance plan, a case-control study was conducted. Patients with at least three mentions of phenome-wide association scans (PheWAS) codes 295 and 295.1 were defined as cases. For each schizophrenic patient, an age-, gender-, and number-of-diagnostic-codes-matched control was identified. For each of the 1,862 disease phenotypes defined by the PheWAS codes, a Chi-square test was performed to evaluate the prevalence rate differences between cases and controls, and the Bonferroni correction was employed to correct for multiple comparisons. The same procedure was used to determine the phenotypical differences between the DSM-IV subtypes of schizophrenia defined by ICD-9 codes, with the unspecified subtypes removed from the analysis. To identify the risk factors of schizophrenia onset, a trajectory analysis was performed by computing the changes in prevalence rates of each psychiatric phenotype within 0-48 months (in 6-month intervals) before the diagnosis of schizophrenia. Wilcoxon signed rank test was used to compare the prevalence rates of psychiatric phenotypes with those in the matched controls.

Results
37,896 schizophrenic patients were identified from the dataset. Bipolar (odds ratio (OR)=21.76, 95% confidence interval (CI)=(20.50,23.09)), anxiety (OR=13.04, 95% CI=(11.87,14.33)), dissociative disorder (OR=9.80, 95% CI=(6.73,14.28)), major depressive disorder (OR=5.46, 95% CI=(5.24,5.68)), posttraumatic stress disorder (OR=5.33, 95% CI=(4.87,5.84)), Parkinson's disease (OR=4.51, 95% CI=(3.95,5.14)), and Alzheimer's disease (OR=4.14, 95% CI=(3.73,4.59)) were significantly more prevalent in the schizophrenic cohort compared with the matched controls (corrected P-value < 0.05). In contrast, schizophrenic patients were less likely to develop malignant neoplasm of retroperitoneum and peritoneum (OR=0.18, 95% CI=(0.10,0.34)) and osteochondropathies (OR=0.24, 95% CI=(0.14,0.39)). Sensitivity analyses showed that the results were not sensitive to the number of PheWAS codes used to define schizophrenic cases. Subtype analysis revealed that 79.30% of the schizophrenic patients have only one subtype based on their ICD-9 codes. Compared to patients with the paranoid subtype, those with schizoaffective subtype are more likely (corrected P-value < 0.05) to have parasomnia (OR=2.72, 95% CI=(1.61,4.58)), obsessive-compulsive disorders (OR=2.06, 95% CI=(1.68,2.52)), and autism (OR=2.06, 95% CI=(1.56,2.72)), in addition to the common manifestations of affective disorders. Trajectory analysis revealed that the prevalence rates of tobacco use disorder and substance abuse increased significantly before the onset of schizophrenia (P-value < 0.01), from less than 7.9% (7-12 months before the diagnosis of schizophrenia) to 16.1% and 19.5% respectively (0-6 months before the diagnosis of schizophrenia). This increase was especially prominent among patients whose disease onset was between 15 and 30 years old.

Discussion
This is a systematic analysis of comorbidities and risk factors of schizophrenia in a nationwide database. Consistent with the literature⁴, we found that bipolar, major depressive disorder, and anxiety are significantly more common in schizophrenic patients. In addition, our analyses revealed less-reported comorbidities, including Alzheimer’s disease, which could point to the neurodegenerative risk in schizophrenic patients. Interestingly, trajectory analysis demonstrated that many young schizophrenic patients suffered from tobacco use disorder and substance abuse in less than one year before the onset of schizophrenia, indicating the potential need for psychiatric evaluation in young adults with issues of tobacco and substance use. Further validation of these findings is needed. Our analysis approach is generalizable to other psychiatric diseases and conditions.

References
Variability in Physicians’ Information Use: Implications for EHR Design

Swaminathan Kandaswamy, MS¹, Patrick Brown, MD², Jenna L Marquard, PhD¹
¹University of Massachusetts, Amherst, MA ²Mission Health, Asheville, North Carolina

Introduction
The amount of information in the EHR is immense, and clinicians struggle to find and synthesize relevant data¹. If EHRs are to support decision-making, we must determine what information is most important to clinicians in various decision-making contexts. Ideally, we could identify a subset of especially important information, and clearly present those to clinicians. This study provides a novel approach to understanding clinician information-needs, moving from assessment of single types of information to sets of information².

Study Methods and Results
Physician participants (20 Faculty, 113 Residents) were presented with brief histories for five patient cases, and associated NICU flowsheet data. For each patient case, participants reviewed the flowsheet data and selected one of 16 diagnoses. They then identified the flowsheet items (amongst 69 items) that influenced their choice. The 69 items were structured into 7 common categories: vitals, physical exam, labs, blood gas, fluids, respiratory, and other. Within each case, physicians who reached the correct diagnosis used very different flowsheet items. Figure 1, shows that for the two cases few items were used by a large majority of participants. Participants also used very different items across patient cases. We conducted a cluster analysis to identify sets of items participants used. Flowsheet items were included in a cluster based on two criteria: 1) greater than 50% (simple majority) of participants chose each item and 2) greater than 75% of participants chose at least one item in the cluster (cumulative agreement). This clustering was validated based on their being physiological indicators relevant to the diagnoses.

Figure 1. Percent of participants who chose each item, with items in clusters (dark gray)

Conclusion
These findings are important in that they show there is not a single set of information a physician could use to identify the correct diagnosis. While adding to design complexity, it may be useful to dynamically tailor information presentation based on the patterns of the individual physician, and on their initial assessment of the patient case.

References
Predicting Appointment No-shows Using Comprehensive Longitudinal Health Information Exchange Data

Suranga N. Kasthurirathne, BEng1,3, Joshua R. Vest, PhD, MPH2,3, Shaun J. Grannis, MD, MS3,4

1Indiana University School of Informatics and Computing, Indianapolis, IN; 2Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN; 3Regenstrief Institute, Center for Biomedical Informatics, Indianapolis, IN; 4Indiana University School of Medicine

Introduction. Missed appointments pose a significant burden to healthcare delivery5,2. Current efforts to predict missed appointments are limited as: (a) they ignore a wide range of behavioral comorbidities, symptoms, previous visit history, appointment wait times, and demographic data4; (b) ignore no-show rates among different types of clinical visits (e.g., primary care, social services, etc.); and (c) do not leverage advances in machine learning for more accurate classification. We seek to address these limitations by using a wider range of clinical, demographic, and behavioral factors to predict appointment no-shows across several specialized care services.

Materials and methods. We analyzed 84,317 adult patients (≥ 18 years old) who had received at least one outpatient visit at Eskenazi Health of Indiana between 2011-2016. We extracted a longitudinal patient record for each patient from the Indiana Network for Patient Care (INPC), a statewide Health Information Exchange (HIE). We identified the four most frequently scheduled appointments types for prediction, and built four data vectors containing tabulated counts of diagnosis identified via ICD codes, previous visit history and patient demographic data. Each data vector was split into randomly selected groups of 90% (training) and 10% (test) data, and used to build a decision model using the random forest classification algorithm, which was selected because of its ability to perform internal feature selection. Each model was tested using the 10% holdout dataset, and evaluated using Area under the ROC curve (AUC) and optimal sensitivity/specificity values calculated using Youden’s J-Index.

Results. The patient population reflected an adult, urban, primary care safety-net population. The most frequently scheduled visit types, together with their AUC, Sensitivity (Sens) and Specificity (Spec) scores were; a) primary care (AUC=0.71%, Sens=0.68%, Spec=0.64%) b) Ob/Gyn (AUC=0.65%, Sens=0.62%, Spec=0.64%) c) dietitian (AUC=0.74%, Sens=0.69%, Spec=0.68%) and d) mental health (AUC=0.73%, Sens=0.72%, Spec=0.64%). Models predicting missed primary care, dietitian and mental health visits outperformed the Ob/Gyn prediction model.

Discussion. Predictions for primary care, dietitian, and mental health appointments yielded results comparable to previous studies3,4. The low performance of the Ob/Gyn model is attributed to the diversity of patients receiving referrals for health issues including birth control, childbirth, and menopause. Our results are more accurate and generalizable than previous efforts as: (a) our patient set covers a large hospital system with a more generalized patient population, and (b) includes comprehensive longitudinal patient dataset representing clinical, behavioral, demographic and visit data. Future work includes predicting other appointment types, integrating aggregate socio-economic factors, and considering temporality and chronological order of clinical events. Though our study population was more generalizable than previous efforts, these results may not be consistent across populations with higher incomes or differing socioeconomic status, and definitions for visit types may differ across health systems and regions. However, this work represents a practical approach that can be applied across different health systems and datasets.

References

An Innovative PHR System for MCH by Constructive Utilization of Infrastructure for Integrating Pediatric Medical Information

Shogo Kato, Ph.D.1, Yoshihiko Morikawa, MD2, Kosuke Nakano, Master1, Takahisa Ogasawara, Bachelor1, Tomoya Ito, MD3, and Naohisa Yahagi, MD, Ph.D.4,5
1National Center for Clinical Research and Development, Tokyo, Japan; 2Tokyo Metropolitan Children’s Medical Center, Tokyo, Japan; 3Aichi Children’s Health and Medical Center, Aichi, Japan; 4Keio University, Kanagawa, Japan; 5The University of Tokyo, Tokyo, Japan

Introduction

There is no boundary between health and medicine in personal healthcare, and it is important to consolidate a variety of life-long healthcare information in order to understand individual health conditions for prevention, diagnosis, and treatment. However, the situation where information is presently scattered across service providers has caused various problems. The Maternal and Child Health (MCH) management system in Japan, that is centered on MCH handbook and aligned with vaccination and infant medical checkup, has been developed in order to provide a series of care for MCH. There are many applications for MCH to manage Personal Health Record (PHR), however, these PHR systems only correct the data from service providers or users in their own forms, and cannot consolidate the data to share it.

The goal of this study is to develop an innovative PHR system that utilizes the infrastructure for integrating pediatric medical information. It will allow us to securely share data between various applications for MCH and local government information systems, based on the condition of the individual consent.

Method

The Clinical Data Management System (CDMS) was implemented as our information logistics technology for integrating health and medical information. CDMS is a user-friendly, versatile, and secure network system for data and knowledge management. CDMS can use various formats to assemble, search, extract and reconstruct different healthcare data from multi-vendor electronic medical records and various healthcare applications. The Personal condition Oriented Proxy System (POPS) was implemented as an infrastructure function to control information distribution based on the condition of the individual consent. It was able to digitalize and consolidate various health and medical information at the individual level.

The dataset required for infant medical checkup was defined and built to extend the existing module of CDMS. Functional improvements were made to the application, which works on CDMS, and the graphical user interface of the application was developed. An environment for transferring and storing data with the outside application was created. The dataset of infant medical checkups, managed by a model local government, was identified among the modules of CDMS. The ability to share information among multiple applications and systems was evaluated.

Results

For 4-month-old infant medical checkups, there were 155 items in the interview sheet and 46 items in the examination sheet. Of those, 80 interview items and 38 examination items were administrated in the document management system by the model local government. These items were mapped for the modules of CDMS. It was confirmed that the ability to provide the information collected via MCH applications for other applications and local government information systems through CDMS based on the state of the individual consent on POPS.

Conclusion

The innovative PHR system for MCH enabled to securely share information generated from the individuals themselves based on the condition of the individual consent, among infant medical checkup applications, vaccination applications, MCH handbook applications, and local government information systems. It was demonstrated that the infrastructure fully utilized the total functions of the MCH management system of Japanese origin.
INTRODUCTION

Current American Heart Association/American Stroke Association (AHA/ASA) Guidelines strongly recommend optimal control of stroke risk factors among ischemic stroke patients. A recent update to these guidelines have outlined best possible care for this patient population, including initiation of a well-coordinated outpatient plan of care while a patient is hospitalized. Implementation of projects designed to improve this transition in care from the inpatient to the outpatient setting is needed.

The goal in developing this Electronic Health Record (EHR) solution is to improve the transition in care for stroke patients by, documenting medication changes, delivering the appropriate education to both patients and families, and ensuring follow-up correspondence and appointments are established.

METHODS

Based on the current AHA/ASA Guidelines regarding post-stroke discharge care coordination, patient/caregiver education, and post-stroke risk factor management, we developed a discharge template embedded within the EHR, to standardize and enhance the delivery of these recommendations. The note created addressed critical transitional areas including hospitalization, medications, education, and follow-up appointments (Figure 1).

The note was developed with input from key stakeholders, including inpatient, specialty care providers, outpatient, primary care providers, nursing staff, clinical pharmacists, and clinical informatics. The note is completed by inpatient clinical staff (e.g., stroke nurse). This note was developed, piloted among 10 neurologists and primary care providers, and revised based upon end-user feedback. Given the complexity in follow-up care, documenting follow-up appointments deemed insufficient to ensure patients received timely post-discharge follow-up. As such, further modifications were made such that patients had all necessary appointments made for them prior to discharge (Figure 2).

RESULTS

This template has been implemented at a single VA medical center as of September 2017. During that time, it has been completed on 15/15 eligible patients. Mean duration to first post-discharge follow-up appointment was 12.2±3.1 days. Implementation of this program also produced pronounced changes in patient care. For example, prior to discharge, 8/13 patients received extended arrhythmia monitoring with another 3 patients receiving it at their first follow-up visit. 2/11 patients had positive findings within 30 days of arrhythmia monitoring, resulting in a change to anticoagulation for stroke prevention.

DISCUSSION

There are critical junctures in the care of stroke survivors where patients experience transitions in healthcare settings (e.g., inpatient to outpatient) and providers assuming primary management of stroke care. In developing an EHR-embedded transition in care note within the Computerized Patient Record System (CPRS) with feedback from end users, stroke survivors are receiving guideline concordant post-stroke care through timelier follow-up, enhanced education, and improved communication between healthcare teams. Longer term follow-up regarding the impact of implementing the transition in care note on the change in the processes of care ordered and its influence on management changes and patient outcomes are necessary. In the short term, assessment of its feasibility, acceptability, and impact of patient flow are planned for spring 2018.
Design considerations for tools supporting multi-centre clinical trials

Omar Khan, MEng1, Sarah N. Lim Choi Keung, PhD1, Siobhan Quenby, MD1,2, Arri Coomarasamy, MD3, Theodoros N. Arvanitis, DPhil1

1Institute of Digital Healthcare, WMG, University of Warwick, Coventry, UK; 2University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK; 3Birmingham Women's and Children's NHS Foundation Trust, Birmingham, UK

Introduction

Miscarriage, defined as the loss of pregnancy before the fetus reaches viability, is the most common complication of pregnancy. As many as 15-25% of pregnancies end in miscarriage1 but, due to the limited data collected, the actual miscarriage rate is unknown. Currently, couples with recurrent miscarriages are routinely screened for various endocrine, immune, anatomical and genetic risk factors2. However, the ability of these tests to stratify women in terms of pregnancy outcome and assign appropriate treatment has not been rigorously tested due to lack of prospective data.

Methods

The Tommy’s National Centre for Miscarriage Research aims to collate information from multiple observational and interventional studies, miscarriage clinics, and maternity databases, across the UK, into one system (Tommy’s Net). It is intended to connect to sites’ clinical IT systems and consolidate information, map data elements between sources, resolve differences in procedures and capture study-specific information for analysis at a national level. Tommy’s Net is based on the CURRe framework3, a modular system for collecting research data in secondary care settings. The framework includes methods for the standardized, flexible capture and storage of data. When developing the system, three network configurations were assessed, by listing the advantages of each and prioritising features in terms of their support for data entry from multiple sites, ease of installation and monitoring, ease of integration with existing systems, their security and ability to collate data and support further development: Distributed – sites store data collected locally, using their own schema and preferred data collection tools. Federated – sites install a local copy of the system and transfer data to a central repository. Centralized – sites enter data directly into a central system.

Results and Conclusion

The Centralized option was chosen as it provides a central database, makes maintenance and recovery of the system simpler and, using a web-based system, allows authorised users to connect from any participating site. One of the participating Trusts was also chosen to host the system, making use of the UK’s secure national broadband network, the Health and Social Care Network (https://digital.nhs.uk/health-social-care-network), and the existing capability of UK trusts to transfer electronic data internally, using HL7 v2 as the communication standard. By exposing a HL7 FHIR interface and connecting to the host trust’s integration engine, the Tommy’s Net system will be able to request data electronically from all participating Trust EHRs in a scalable manner as long as the trust IT teams allow access from the host trust. Other integration mechanisms will be considered based on the standards adopted by the Trusts.

The Tommy’s Net system is currently running at UHCW and is being used by research midwives to enter miscarriage data. Other participating sites will start using Tommy’s Net in the coming months and approval has been obtained to integrate Tommy’s Net with Trust EHRs to share demographic and other information for patients registered with the Trust, avoiding redundant data entry by allowing data already collected by the Trusts to be imported. Tommy’s Net has been developed for use by miscarriage services within the UK to enable population-based epidemiological studies and facilitate large cohort studies. The design of the system promotes interoperability with existing Trust systems to allow researchers to use information already collected, and collect pregnancy outcomes, so that participating Trusts can benchmark clinic success rates and identify high risk groups of patients for future research.

References

Matched Pair Network Analysis of Chronic Condition Predisposing to Depression

Min-hyung Kim, MD, MS,1 Samprit Banerjee, PhD,1 Yize Zhao, PhD,1 Fei Wang, PhD,1 Yiyi Zhang, PhD,1 Yongjun Zhu, PhD,1 Joseph DeFerio, MPH,1 Sang Min Park, MD, MPH, PhD,2* Jyotishman Pathak, PhD1*

1Joan & Sanford I. Weill Medical College of Cornell University, New York, NY, USA; 2Seoul National University College of Medicine, Seoul, Korea

* Correspondence to: Sang Min Park <smpark.snuh@gmail.com> and Jyotishman Pathak <jyp2001@med.cornell.edu>

Introduction

Although the literature has abundant evidence for correlation between depression and chronic physical illness, comprehensive understanding of the complex relationships between multiple chronic conditions and depression requires a randomized sample at scale, beyond convenient, selective and specific study populations in traditional studies1,2. In this study, we present a network-theoretic analysis to explore all possible pairs of chronic medical conditions that differentially co-occur in newly diagnosed major depressive disorder (MDD) patients compared to their matched controls from Korean National Health Insurance Services3.

Methods

Newly diagnosed MDD patients were matched to controls for demographic profiles, socioeconomic status, health insurance status, and geographical locations in Korean National Health Insurance Services, using the scale-invariant Mahalanobis distance-based nearest neighbor algorithm4. To systematically compare the network of chronic medical conditions in the case groups of MDD and the controls after adjusting for confounding factors, we performed 2-sided test under the null hypothesis that the Fisher-transformed difference between the correlations in cases and controls equals to zero5. P-values for rejecting the null hypothesis of Fisher-transformed correlations being the same for cases and controls were calculated under the null permutation distribution by shuffling the case/control labels within each matched pair.

Results

With the matched pair network analysis, we prioritized pairs of chronic medical conditions which show differential co-occurrence in depressed individuals and controls. In particular, the matched pair network analysis prioritized ischemic heart disease and obstructive lung disease, chronic headache and chronic ophthalmologic disorder, epilepsy and chronic headache, epilepsy and chronic neurologic sequelae, diabetes mellitus and chronic headache. The prioritized pairs of chronic medical conditions showed statistically significant variable interaction in the generalized linear (log-odds) scale in response to the probability of being diagnosed of MDD.

Discussion

Matched pair network analysis provides a systematic view of the differences in the disease networks in in newly diagnosed MDD patients compared to their matched controls, after adjusting for demographic profiles, socioeconomic status, health insurance status, and geographical locations. The network analysis of this study identifies potential pairs of chronic medical conditions which show differential co-occurrence in depressed individuals and controls, and these pairs could be prioritized for further investigation.

References

Personalized Patient Education for Systemic Vasculitis
Sujin Kim, Ph.D.,1, Aleksander Lenert, M.D., M.S., FRCPC1, Zhenhui Piao, M.S.,1, Anil R. Kumar, M.S.,1, Qiang Cheng, Ph.D.1 University of Kentucky, Lexington, KY, USA

Introduction

Systemic vasculitis are a rare group of immune-mediated disease that result in inflammation of blood vessels and resultant ischemia with subsequent end-organ damage and life threatening complications. As SV is rare and its etiology remains elusive, it is challenging for patients with vasculitis to adequately grasp their condition, monitor symptoms and understand complex treatment. During rheumatology clinic visits, patients are provided with education pamphlets (patient education materials or PEMs) related to their disease; however, these PEMs are not personalized, lengthy and written well above the recommended readability level for our patients.

Methods

In order for us to design personalized PEMs for vasculitis patients, we performed multiple informatics techniques such as data mining, clinical natural language processing, and bibliometric visualization. The following describes further details about three techniques we applied and study data sets we used in our analysis. Our study intends to answer the following two key research questions: (1) patterns of any associations between vasculitis clinical visits in terms of services used; and (2) identification of key concepts contained in professional literature and PEMs.

We extracted clinic visit histories and service details from the i3 InVision Data Mart database and medical literature data set for each of four SV types were retrieved from the PubMed database which include 6,725 articles for PAN, 8,311 for GPA, 6,795 for GCA, and 4,299 for TAK. For Frequent Item Set Mining, we used Frequent Pattern growth (FP-Growth) using pyfpgrowth 1.0 in Python package. For detection of major concepts in the major vasculitis PEMs, we used a clinical natural processing (cNLP) application called Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP) 1. We used the VOSViewer to construct major concepts and their relations among vasculitis journal articles from PubMed.

Results

RQ1: Patterns of Clinical Visits among Vasculitis Patients: Based on all four observations of FP-Growth results, we found the following observations. Management of all four types of SV (TAK, GCA, PAN and GPA) involves an outpatient clinic visit, venipuncture (i.e. blood draw) and assessment of specific laboratory values for diagnosis, monitoring of disease activity and medication side-effects. Additionally, markers of inflammation, specifically ESR (erythrocyte sedimentation rate) and CRP (C-reactive protein), are routinely checked and monitored in all four types of SV. Additionally, CBC (i.e. complete blood count) is monitored (or ordered) for all four types of SV.

RQ2: Patterns of Major Concepts in Medical Literature by Four Vasculitis Conditions
A total of 6 clusters were displayed in Figure 1 which contains 28,856 links and total link strength is 84,478 based on minimum cluster size 20. In Figure 1, the yellow colored group represents basic science such as immunoglobulin and pathophysiology issues while the aqua colored group represents kidney related concepts such as glomerulonephritis and renal hypertension.

Conclusion

Patients diagnosed with SV are confronted with a number of diagnostic and therapeutic procedures. Thus, vasculitis patients have to actively engage in their care by understanding complex medical instructions during the initial diagnosis of their rare disease 2. Generic PEMs on SVs are not customized to individual patients whose information requirements differ depending on their disease course.

References

Mapping data items between EDC and EMR using FHIR terminology service

Eizen Kimura, B.M, Ph.D. 1, Keiichi Yamamoto, Ph.D.2
1Medical School of Ehime University, 2Osaka City University Graduate School of Medicine

Abstract

We annotated Operation Data Model (ODM) files using standard terminology and employed the FHIR terminology Web service to map between local and standard terminologies. This enabled efficient matching of data items between Electronic Data Capture (EDC) and Electronic Medical Record (EMR) systems.

Introduction

REDCap, developed by Vanderbilt University, is Electronic Data Capture (EDC) project supporting the Clinical Data Interchange Standards Consortium (CDISC) Operation Data Model (ODM). Previous study described automated mapping of the ODM of the REDCap instrument to a Study Data Tabulation Model (SDTM) 1. However, individual vendors or healthcare providers must link EMR data that are not collected in clinical studies when creating EDC research data collections. Thus, mapping between EMR and ODM terminologies is required. To this end, we annotated ODM files using standard terminologies and developed a mapping tool using FHIR terminology service 3.

Methods

We modified REDCap to allow importation of the ‘Alias’ child components of the ‘ItemDef’ XML elements used to annotate semantic definitions in the ODM; we placed these contents in the REDCap annotation fields. To automate mapping of EMR items to REDCap fields, we developed a Web service that publishes EMR metadata; the service is compliant with the Dynamic Data Pull (DDP) interface used by REDCap to import external data. The FHIR Terminology service (FHIR TS) was configured to map the concepts of both the JLAC10 (Japan Laboratory Code Ver.10) and the UMLS (Unified Medical Language System) (Fig. 1) because, in Japan, laboratory data are usually managed (stored in EMRs) employing the JLAC10 code system.

Result

We imported an ODM-based questionnaire file in which every item was annotated with UMLS concepts into REDCap. Prior to using the questionnaire, we changed certain configurations. First, laboratory items annotated with UMLS concepts in the questionnaire were extracted from the ODM file. Then, the FHIR TS was employed to translate the UMLS concepts to those of JLAC10. Our new tool matches the translated JLAC10 codes to the JLAC10 codes of EMRs available in the Metadata Service of the DDP interface. Finally, the tool populated a “redcap_ddp_mapping” Table and EMR data automatically imported by REDCap through the Data Service of the DDP interface. We confirmed that REDCap placed EMR data in the appropriate fields of the questionnaire.

Conclusion

Annotation of ODM files with standardized terminology followed by leveraging of the FHIR terminology service enabled efficient mapping of data items between EMRs and EDCs. Currently, every EMR vendor must implement a unique Web service containing the metadata of all data items in EMRs because the information models are vendor-specific. In future, we will try to develop a fully automated mapping information model using EMR terminologies.

References

Understanding Factors Behind Diagnostic Pitfalls Through Quantification Of Diagnostic Heuristics

Mikhail Kovalenko, MS1,2, Richard Hammer, MD1, Dmitriy Shin, PhD1,2
1Department of Pathology and Anatomical Sciences, University of Missouri, Columbia, Missouri; 2MU Informatics Institute, University of Missouri, Columbia, Missouri

Abstract

Diagnostic decisions are a result of a complex mental activity that is prone to diagnostic pitfalls. Through the use of activity capturing tools, informatics algorithms, and descriptive measures, we attempt to understand and quantify the steps, heuristics and results of the diagnostic process so that it can be explained and improved upon.

Introduction

Behind lymphoma diagnosis there is a complex mental process influenced by a multitude of factors ranging from pathologist’s experience to circumstances of the clinical case that sometimes results in an incorrect diagnosis due to diagnostic pitfalls. While many pitfalls are well known, some may not be obvious. In order to understand the reasons for pitfalls and their effect, we use whole-slide imaging (WSI) and informatics tools to reveal and quantify the mental and visual heuristics that contribute to diagnostic decisions.

Materials and Methods

Using our PathEdEx whole-slide-imaging platform, we collect pathologists’ activities, diagnostic clues, and their mental and visual heuristics associated with diagnosing cancerous tissue slides within a context of a complete pathology case. To quantify the highly heterogeneous data for subsequent processing and analysis, we arrange the unorganized collection of actions, clues, and heuristics into ordered sequences and vocabulary-based datasets that are suitable for association rule mining (ARM). ARM-induced association rules are then filtered to retain only those related to diagnostic outcomes. For each diagnosis, the rules can be compared on the basis of their content, rule confidence, and lift. The information gain of each rule \( Y \) is calculated using Kullback-Leibler (KL) divergence (Eq. 1) to show its contribution to the diagnosis in comparison to other rules \( X \).

\[
D_{KL}(Y||X) = \sum_{i=1}^{N} \ln \left( \frac{Y_i}{X_i} \right) Y_i
\]

Equation 1.

Results and Discussion

The initial dataset was produced from eight pathologists diagnosing eight anonymized lymphoma cases. Our analysis shows that the likelihood of a correct decision may increase threefold (rule confidence 30%→100%) with the addition of new heuristic items, one at a time, to an existing diagnostic itemset. Useful additions result in information gain quantified by a greater-than-zero KL divergence \( D_{KL}>0 \). Once the likelihood peaks, additional diagnostic clues do not contribute any more information \( D_{KL}=0 \). KL divergence allows us to find the smallest set of diagnostic heuristics that maximizes the likelihood of the correct diagnosis for a given phenotype, providing opportunities for diagnostic process optimization. Minimization of KL divergence between itemsets associated with incorrect decisions and itemsets that point to correct diagnoses enables the identification of borderline heuristics that may lead to diagnostic mistakes. The method’s descriptive nature may not be directly useful in prediction of new cancers.

Conclusion

We see the main application of this method in improving pathology education through development of optimum decision paths leading to unambiguous diagnoses. We believe that a clearer understanding of pitfall-contributing factors and discovery of more efficient diagnostic pathways will prompt practitioners to take conscious steps towards avoidance of potential pitfalls during the diagnostic process.

References

Registry on FHIR: Leveraging HL7’s Newest Standard for High Quality, Multipurpose Clinical Data Repositories

Philip Langthorne, Patrick Young, PhD, Leon Rozenblit, JD, PhD
Prometheus Research, LLC, New Haven, CT

Introduction
In the past, clinical data registries focused on either clinical research or on quality improvement. More recently, resource constraints and increasing need for actionable, repurposable data have inflated the demand for registries that can support both use cases equally-well, requiring unprecedented scalability, data quality, security, and agility. Data acquisition methods have typically included document exchange, creation of registry-specific information exchange schemas, manual chart abstraction, and creation of one-to-one mappings from individual data suppliers into the registry. For instance, the Informatics for Integrating Biology and the Bedside (i2b2) approach calls for developing and maintaining detailed mapping tables (Klann et al., 2016), which requires deploying significant resources with these costs borne by all participants in the registry network. We determined that no existing approaches reached the cost-effectiveness threshold we were seeking and decided to pursue a new one, with the emergent Health Level Seven Fast Healthcare Interoperability Resources (HL7 FHIR) standard at its core.

Method
We developed an application programming interface (API) using the FHIR Draft Standard for Trial Use 2 to allow Electronic Health Record (EHR) vendors to submit data directly to the registry using standard web technologies. The API is written in Python and enables system-to-system interactions similar to SMART on FHIR (Mandel et al., 2016), except that the SMART on FHIR model is designed primarily for contained applications whereas our approach is designed primarily for bulk data transfer. The API uses HTTPS with certificate-based authentication to ensure security, and provides rigorous, real-time validation of the data received. To more efficiently distribute development costs, the API is also concurrently deployed to a sandbox environment, so that data suppliers can control their own cost curves as they work on integration. FHIR objects that pass validation are stored in a clinical data warehouse built on PostgreSQL using an entity-attribute-value schema design, where the “value” is the FHIR object itself. Finally, to support flexible applications of the data, the FHIR objects are streamed into purpose-specific data marts, where they are renormalized into each mart’s information model for analysis activities, thus ensuring that no one use of the registry restricts simultaneous or future data repurposing for technical reasons.

Results and Discussion
This approach has allowed a project team of five to engage 10 cloud-based EHRs in the first year of the registry. Furthermore, we found the FHIR standard to be highly accessible and were able to go from never having used FHIR before to releasing our API to production within three months. The granularity of the standard helps assure the quality and agility of the registry data. It also permits EHRs to only transfer changes to a patient’s record, reducing latency relative to document-based and manual approaches. Within the registry warehouse, our data modeling approach further differs from i2b2 in that we employed a higher level of normalization, containing all clinical facts on a single table to simplify retrieval, whereas i2b2 distributes information across multiple dimension tables. Finally, although we have observed a slower than desired uptake of the FHIR standard by data providers, we note the enthusiasm and growing consensus around the standard and we anticipate further cost reductions to accrue to all parties in the registry program as the standard continues to mature and introduce its first normative releases.

References
Online Support Group as Means to Improve Oral Health Literacy for People with Temporomandibular Disorder
Karen Lin, BS, Joseph Finkelstein, MD, PhD
Columbia University, New York, NY

Introduction: Temporomandibular Disorder (TMD) is characterized by jaw joint dysfunction which is manifested by pain in the jaw joint, facial structures, head, and muscles controlling the joint movement. The exact causes and cure of TMD remain unknown. As a result, individuals affected by TMD continue to experience high amount of tension and anxiety, and frequently use online support as means to improve their health literacy. The aim of this study was to examine TMD health literacy issues and solutions addressed in online communities by identifying the pattern, accuracy and reliability of the information exchanged on online support group among TMD patients who sought empowerment, social support, and informational gain.

Method: An online TMD support group was identified for the purpose of this study based on high quality, popularity, topic diversity and responsiveness (http://www.mdjunction.com/forums/tmj-discussions). Online postings were identified on the basis of a minimum of 5 messages per thread; diversity in personal experience, emotional similarities and differences, and informational knowledge. The posts selected for further analysis and study were chosen based on the types of topic (story-telling, advice seeking, emotional distortion, etc.), the amount of information shared, the level of participation and responsiveness from the members, and values of the context. A final analytical sample of 121 messages has been used to allocate and identify pattern and themes that were found in the data until saturation has been achieved. Messages were grouped into four main categories by using six identification themes developed for this study: M for Money - the cost of care and treatment for TMD patients, P for Pattern - individual pattern of symptoms and relations, PL for Personal Life - stories of treatment or discomfort, E for Empathy - expressing and expecting empathy towards members, EX for Exchange - exchange of informative resources and knowledge, and MD for Medical - clinical treatments, informative knowledge, medical resources, and healthcare.

Results: Emotional issues accounted for 38.8% (n=47), informational issues - 33.8% (n=41), counseling issues - 15.7 (n=19), and cost of care issues - 11.5% (n=14), respectively, of the exchanged context. Online group participants connected to each other by expressing sympathy or empathy toward the writer of the post. By providing similar views and personal experience, members were validating and acknowledging each other’s symptoms and pain. Not only emotional and mental support but also health literacy exchanges helped the patients be empowered and engaged in their self-care. Providing informational support also appeared to be extremely helpful to members, who encouraged each other to become informative about TMD. Online support group provided a platform where people could easily obtain beneficial feedback regarding a successful or failed treatment. Due to the low success rate of TMD surgery and treatment, members expressed fear and sought advice from other members regarding their past experience and personal life, in addition to maintaining a positive and hopeful tone through encouragement. However, one of the least mentioned yet impactful topic discussed among patients showed frustration towards disorder or related disease, for which they tended not to have enough ‘savings’ to acquire the best treatment they need. Online group participants also exchanged information about their work benefits and insurance coverage which exposed the healthcare issues germane to TMD cases.

Conclusion: The significant proportion of emotional and informational support indicated that online interaction facilitated patient empowerment by providing essential components for TMD self-management that were not made available sufficiently via alternatives venues. Online support groups may successfully supplement other interactive tools such as computer-assisted education and patient portals to enhance oral health literacy and facilitate patient engagement in self-care.
Comparisons of Prescribing and Dispensing Data on Medications in the PCORnet Antibiotics and Childhood Growth Study

Pi-I D. Lin, MS, ScD1, Sheryl L. Rifas-Shiman, MPH1, L. Charles Bailey, MD, PhD2, Janne Boone-Heinonen, PhD, MPH3, Matthew F. Daley, MD4, Christopher B. Forrest, MD, PhD2, Casie Horgan, MPH1, Darren Toh, ScD1, Jessica Young, PhD1, Jason P. Block, MD, MPH1

1Harvard Pilgrim Health Care Institute, Boston, MA; 2Children’s Hospital of Philadelphia, Philadelphia, PA; 3Oregon Health & Science University, Portland, OR; 4Kaiser Permanente Colorado, Denver, CO

Introduction

Objective information on prescription medication utilization requires either electronic health record (EHR) prescribing records or dispensing data from pharmacies or insurance claims. Each source has limitations. For example, prescribing records will miss prescriptions given outside of the health system providing data, and dispensing records from insurance claims will miss prescriptions paid for out of pocket. Comparing prescribing-dispensing records enhances the understanding of medication exposure, helps determine the rate of unfilled prescriptions, and provides insight into the underlying mechanism of medication utilization. Using data from the National Patient-Centered Clinical Research Network (PCORnet) Antibiotics and Childhood Growth Study, we assess the linkage between prescribing and dispensing data among children from birth to 11 years of age.

Method

Twenty-eight PCORnet Network Partners (NPs) participated in the Antibiotics and Childhood Growth Study. Among eight NPs with both prescribing and dispensing data in the PCORnet Common Data Model (CDM), prescribing and dispensing data were linked bi-directionally using script- and episode- approaches. Same-day scripts were de-duplicated, giving priority to broad spectrum antibiotic scripts; one script within a 10-day window was kept for the episode approach. We examined different linkage methods using patient ID, time between records (0-7 lag days), and antibiotic classification. When more than one linkage was present, we kept the linkage best matched by antibiotic class. We included only oral antibiotics scripts from ambulatory visits and stratified by integrated delivery system (IDS) and non-IDS. We used logistic mixed effects regression, accounting for clustering by NP, to examine clinical and demographic variables associated with having a linked script.

Result

We achieved the highest linkage rate by using patient ID, antibiotics spectrum, class and sub-classification, 7-lag days, and the script approach. Under these parameters, we linked 91.2% of prescribing (N=257,779) and 95.1% of dispensing (N=238,526) records from IDS sites and 61.4% of prescribing (N=161,305) and 78.2% of dispensing (N=122,770) records from non-IDS sites. We found similar results when using the episode approach. Compared to broad-spectrum antibiotics, narrow-spectrum antibiotics had higher odds of linkage in both directions. Among commonly used antibiotics, penicillin, 3rd generation cephalosporin, and penicillin combination medications had higher odds of linkage while macrolides and sulfa had lower odds of linkage. Other factors associated with higher linkage included age, race, type of diagnosis, gender, and sites. For example, compared to prescriptions recorded for children ages 0-6 months, those for ages 60-132 months had higher odds of linkage (OR=1.15, 95% CI=1.06-1.23 for IDS and OR=1.77, 95% CI=1.67-1.88 for non-IDS comparing 60-132 months to 0-6 months). Compared to prescriptions associated with diagnoses for which antibiotics are always indicated (e.g. pneumonia), prescriptions associated with diagnoses not requiring antibiotics (e.g., nonsuppurative otitis media) had lower odds of linkage (OR=0.53, 95% CI=0.50-0.56 for IDS and OR=0.84, 95% CI= 0.80-0.86 for non-IDS). Dispensing records from females (OR=1.23, 95% CI=1.19-1.28 for IDS and OR=1.07, 95% CI=1.04-1.10 for non-IDS) had higher odds of linkage compared to male while those from Blacks or African Americans (OR=0.78, 95% CI=0.75-0.84 for IDS and OR=0.85, 95% CI=0.82-0.88 for non-IDS) had lower odds of linkage compared to Whites.

Conclusion

Integrated delivery systems had much higher linkage rates between prescribing and dispensing records. Some differences in overall matching rates were evident by antibiotic class. Quantifying the linkage pattern between prescribing and dispensing data can be used to develop an algorithm to more accurately define antibiotic use among the pediatric population, especially by further examining individual-level and system-level predictors of missing data.

References

City-wide Opioid Use Among Patients Treated at Multiple Health Care Facilities

Margaret B. Madden¹, Marc B. Rosenman, MD¹,²

¹Center for Health Information Partnerships, Northwestern University, Chicago, Illinois; ²Department of Pediatrics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois

Abstract
Opioid abuse represents a significant problem in the United States, and there is a growing awareness that fragmentation of medical information poses a challenge to patient safety. Using a de-identified¹ assembly (HealthLNK) of electronic health records from seven institutions, we sought to quantify, at the population level, the burden of fragmentation in opioid use across the city of Chicago by measuring the rates at which individuals are prescribed opioids across multiple sites.

Study Design
In a retrospective (2006-2012) cohort study of five large healthcare institutions from HealthLNK, we analyzed the percentage of primary care patients (Chicago address and at least two visits) with an opioid prescription, stratified by the number of institutions the patient visited overall. We then describe the principal diagnoses at visits where an opioid was prescribed, stratified by number of institutions where the patient received an opioid prescription, and by the overall number of institutions the patient visited. We also examined number of emergency department (ED) visits per patient, stratified by overall number of institutions the patient visited.

Principal Findings
Among 884,438 primary care patients in the 7-year period, 455,400 (51%) received an opioid prescription. Among the 884,438, about 25% (209,419) had visited ≥2 institutions; 3.5% had visited ≥3 institutions, and 0.3% had visited ≥4 institutions. As the number of institutions per patient increased, the percentage of patients receiving opioids also increased (Figure 1).

Figure 1. Primary care patients with an opioid prescription

Pain-associated chief complaints (chest, abdominal, joint, back, limb, neck, or dental pain, headache, or pain not elsewhere classified) at opioid Rx visits were 3-fold more frequent among the 5-institution patients than among the 1-institution patients. The percentage of patients with an ED visit also varied directly with number of institutions visited: 25%, 45%, 62%, 88%, and 93% in the 1, 2, 3, 4, and 5-institution groups, respectively.

Conclusion
The percentage of patients who have received opioids increases with the number of institutions visited. Patients at risk for drug-seeking behavior may have higher city-wide visit rates. Of course, patients who truly have more severe or refractory illnesses also may be more likely to visit multiple institutions. This report is a prelude to further analyses that control for confounders and examine additional data sets. With the current need for a solution to the opioid epidemic, improving communication and integrating health records between sites may help clinicians identify patients who are taking advantage of the lack of information associated with fragmentation of care.

References
A FHIR-compliant approach for exchanging data between Decision Support Systems for diagnosis and Electronic Health Record systems

Samhar Mahmoud\textsuperscript{1}, Derek Corrigan\textsuperscript{2}, Brendan Delaney\textsuperscript{3}, Vasa Curcin\textsuperscript{1}
\textsuperscript{1}King’s College London, London, UK; \textsuperscript{2}Royal College of Surgeons in Ireland, Dublin, Ireland \textsuperscript{3}Imperial College London, London, UK;

Introduction

The Learning Health System (LHS) paradigm relies on routine extraction, aggregation and transformation of medical data from a variety of sources into actionable clinical knowledge. EHRs are a prime source of data for the LHS in pattern discovery, clinical trial recruitment, and diagnostic decision support systems (DDSS). The acceptance of DDSS has been hampered by a number of challenges, mostly related to usability and added workload for the clinician. A significant challenge identified is the need for fully integrated DDSS into EHRs. Before addressing semantic integration and data privacy issues involved in the communication between a DDSS and an EHR, there is a need to agree on the messages exchanged in the process of performing a recommendation. In response, we present an abstract architecture outlining the main components of a DSS system, and we outline the messages and data content required at each step of the diagnostic supported consultation that selects between differential diagnoses. The architecture has been implemented in the EU FP7 TRANSFoRm project\footnote{https://www.transformproject.eu} and integrated with leading UK primary care EHR systems. HL7’s FHIR specification\footnote{https://www.hl7.org/fhir/} also aims to facilitate the exchange of data between healthcare applications, through the use of standardised APIs. Therefore, the protocol we present in this work is consistent with the FHIR approach, and could be abstracted as a FHIR service.

Methods: Abstract Integration Model

We assume a general DSS is split into three logical units. Evidence Service (ES): the diagnostic knowledge base; Decision Support Mediator (DSM): coordinates communication between the EHR and ES; Decision Support Interface (DSI): a graphical front-end embedded into the EHR. The sequence of interactions comprises three phases:

1. **Initialisation and Data Extraction:** The diagnostic consultation starts by extracting patient EHR data, while the DSI capture the main Reason for Encounter (RfE). All data is then passed to the ES as a diagnostic question, secured according to the recommendations of the FHIR standard. The ES response is an initial ranked list of diagnoses to consider, each accompanied by a list of cues and examinations pertinent to each diagnosis.

2. **Data capture:** Further diagnostic cues are captured in a structured manner by the DSI. In each iteration, every newly captured cue is sent to the DSM and the ES to obtain an updated ranked differential diagnosis list for display to the clinician, followed by an optional capture of a working diagnosis.

3. **Data Storage:** The final step is to write back the captured diagnostic cue data into the patient record using an EHR compatible format. The expected adoption of FHIR as a standard across commercial EHRs will provide additional support for this step.

Discussion and Conclusion

This work presents a first step towards the standardisation of the integration between decision support tools and electronic health record systems. We have outlined the main interactions steps that are needed to perform such an integration and ensured that our approach is compatible with FHIR and offers a means of managing the strict requirements of unambiguous and structured clinical concepts in an environment where terminologies are both widespread and frequently applied by users.
Cross Network Directory Service: Enabling Meaningful Collaboration Across Organizations

Jessica M. Malenfant, MPH\(^1\), Jenny Hochstadt, MS\(^1\), Kimberly Barrett, MPH\(^1\), Zachary Wyner, MPH\(^1\), Daniel Dee, MS\(^2\), Dean Corriveau\(^2\), Bridget Nolan\(^1\), Chayim Herzig-Marx, PhD\(^1\), Jeffrey Brown, PhD\(^1\)

\(^1\)Harvard Pilgrim Health Care Institute, Boston, MA, \(^2\)Avacoda LLC, Amherst, MA

**Introduction:** With the proliferation of electronic health records from EHRs and claims systems, data are becoming more available for research and other secondary purposes. The diversity of electronic health data sources creates challenges to identify data resources and potential collaborators that would enable a true learning health system.

We will describe implementation approaches and challenges creating a platform, the Cross Network Directory Service (CNDS), for researchers and data holders to securely and efficiently share information about their expertise, data sources, and technology resources in accordance with local governance including IRB policies. The focus of this work was to extend the query health application, PopMedNet (PMN), the leading tool used in large-scale distributed healthcare research, to address these challenges.

Prior work in this area has focused on discovery of scientific resources and biospecimens and systems to support professional networking. These projects helped inform our approach to designing and implementing a system for multi-site collaboration, which will lead to opportunities for data source characterization.

**Methods:** The Cross Network Directory Service system includes 4 components: Governance, Registration, Discovery, and Communication. The project had 3 workstreams: (1) develop a data model that includes the following components: CNDS entities (user, organization, data source), metadata domains that change and vary by entity, and visibility governance. (2) Design and develop flexible software, based on PopMedNet, that includes data model-driven code generation (e.g. user interface changes dynamically based on the data updated in the database) without disruption to the end-users. (3) Implement technology that seamlessly respects the governance of multiple PMN instances (e.g. Sentinel and PCORNet) in user “registration” in CNDS for these network members and in cross-network communications (e.g. sending a query across the networks).

This project did not focus on determining which data or metadata should be collected for the various entities. Instead, we prioritized creating a system with extreme flexibility as metadata standards and inventories will grow over time, and forcing hard-coded database and software changes is not scalable or reasonable. The user interface includes metadata management modules and profile pages for users to update information about themselves, their organizations, and their data sources, including the visibility governance related to who they want to share their information with; each screen is dynamically updated according to the metadata management tool that updates the underlying data model.

The work also included developing tools that allow users to search for entities based on metadata concepts and visibility settings; this Discovery module then drives the Communication process that enables a user to send data requests to the data sources in a secure environment using standard PMN query functionality.

**Results:** The data model has been successfully integrated with the CNDS infrastructure that uses PMN version 6.0. The metadata management module, search functionality, and cross-network query capability have been implemented and validated in the CNDS test environments. We have demonstrated the ability to send a data query from a CNDS Sentinel network to the PCORNet network without disrupting existing workflows, while adhering to the governance policies for both networks. Specifically, the system is being implemented in a way that maps request types and electronic workflows between the PCORNet and Sentinel PMN instances. This enables each network both to distinguish cross-network data queries from the within network queries to which they are contractually committed and to customize workflows and related software-enabled governance for cross-network queries.

**Discussion:** Challenges to be discussed include striking a balance between developing a very flexible data model and the impacts of fully implementing it as a dynamic application interface within PMN and promoting metadata curation. Future work includes incorporating this proof-of-concept scaffolding into the current PMN codebase.
Achieving Single IRB Review via the SMART IRB Online Reliance System

Anupama Maram, MS¹, Katherine K Lau, MS¹, Carol A Pech, PhD², Maria A Cervone, MS¹
¹Harvard Medical School, Boston, MA; ²University of Wisconsin-Madison, Madison, WI

Introduction

Multi-site studies have traditionally required review by the Institutional Review Board (IRB) at each study site, incurring a significant burden to the investigator and lengthy lead-times to initiate research, obtain scientific results, and ultimately address patient needs. Over the past decade, regional IRB reliance networks have emerged to reduce duplicative reviews while ensuring a high level of protection for research participants; these networks consolidate IRB review by designating appropriate “reviewing IRBs” on a study-by-study basis, reducing the burden of review across the network. Recently, the National Institutes of Health (NIH) issued a Single IRB Review policy for NIH funded non-exempt human subject research grants or contracts, effective Jan 25, 2018, and established SMART IRB to provide a roadmap for implementing this policy. The core of the SMART IRB initiative is the Master Common Reciprocal IRB Authorization Agreement—a “treaty” that allows one institution to conduct IRB Review on behalf of another, eliminating the need to negotiate agreements on a per study basis. Regional Ambassadors aid in the adoption and implementation of the Agreement, as do procedures, checklists, and other resources. All 64 Clinical & Translational Science Award (CTSA) Hubs have joined the Agreement, and over 350 institutions are currently participating. The SMART IRB Online Reliance System, launched in May 2017, provides a national platform for investigators to request IRB reliance, and for institutions engaged in any human subjects research to collaboratively determine who will review and who will rely via a transparent and automated workflow designed to engender trust.

Discussion

Through system managed workflows and automated notifications, reliance requests are handed off from one user to another as tasks are completed along a path that culminates in the issuance of a determination letter (Figure 1). The letter documents the reliance arrangement, and emphasizes the study team’s responsibilities for next steps. In addition, the system provides a summary dashboard that contains context-sensitive information based on the request’s status and the user’s role, so that each party involved in the request understands the exact status of the request as it makes its way through the workflow: what steps are completed, remaining steps, who currently has an action to take, and who to contact if you have questions.

User research and usability studies were conducted to ensure the interface allows both novice and expert users to navigate the system effectively.

Conclusion

The SMART IRB Online Reliance System fulfills a unique informatics niche to support and standardize IRB reliance decisions at a national scale as NIH implements its new Single IRB Review Policy, and is backed by deep experience in IRB reliance at Harvard, Wisconsin, and institutions across the country. This system has played a critical role in advancing collaborative research across the country by facilitating the reliance review process.

References


Funded by the NIH National Center for Advancing Translational Sciences, grant number UL1TR001102-04S1.
The Risk of Patient Re-Identification Based on Unique Groups of ICD Codes

Elizabeth N. Mulvey, JD¹, Jihad S. Obeid, MD¹
¹Medical University of South Carolina, Charleston, SC

Introduction
International Classification of Diseases (ICD) diagnosis codes are not considered by HIPAA to be among the 18 patient identifiers, and are therefore widely distributed as part of de-identified data sets. In a small data set of 2762 patients, Loukides et al. reported that more than 96% were uniquely identifiable by their combination of diagnosis codes¹. We set out to determine whether these results could be replicated in a much larger sample of patients from the Medical University of South Carolina (MUSC) population and to assess the magnitude of the risk of re-identification for MUSC patients based on ICD-9-CM and the newer ICD-10-CM codes. We also examined patient characteristics that increased the likelihood that a patient’s combination of diagnosis codes would be unique.

Methods
We obtained de-identified patient data set per HIPAA standards, including age (rounded down to years), gender, race, encounter type, and diagnosis codes, for all patients who were seen at MUSC during the calendar year 2015. The population included 194,967 unique patients ranging in age from newborn to over 85, with a total of 686,573 patient encounters, comprised of inpatient, outpatient and emergency department encounters. We then extracted a list of diagnosis codes for each patient (sorted and de-duplicated) and compared the lists of diagnosis codes to identify patients who shared an exact combination of codes with at least one other patient. Based on the results, we classified each patient’s data as “secure,” if the combination of diagnosis codes was shared with at least one other patient, or “unique” or “at risk,” if the patient’s combination of codes was not replicated in the data set.

Results & Discussion
Our analysis showed that 138,314 (73%) of patients in the study population were uniquely identifiable by their combination of diagnosis codes (Figure 1). The number of codes attributable to a patient was the single strongest predictor of unique status, followed by the occurrence of an inpatient admission. Gender and race were not found to be predictors of at-risk status. Our analysis demonstrates that there is a significant risk that patients might be re-identified by their combination of diagnosis codes. The overall percentage of MUSC patients who are at risk is somewhat lower than that previously reported by Loukides (73% vs. 98%). However, for certain groups of patients selected by age, number of codes and type of encounter, our results are consistent what was previously reported. The finding that a substantial percentage of patients can be uniquely identified even as the number of patients grows very large is concerning, as it reduces the likelihood that the use of larger data sets would be an effective strategy to prevent re-identification. Our study identifies particular subgroups of patients who are at highest risk. Furthermore, 71.5% of patients whose data was classified as secure were rendered unique when age, race and gender were also considered. Of note is that Loukides’ report preceded the introduction of the more granular ICD-10 in 2015, which would significantly increase risk of re-identification. An incidental finding warranting further investigation was that 4464 patients (2.2%) had one or more diagnosis codes that were unique to them, raising a different means of re-identification.

Conclusion
Unique combinations of ICD codes may serve as fingerprints for individuals in electronic health records. However, the risk of re-identification is only manifested when combining this data with non-research clinical operational data that includes patient identifiers, and in this case, the “fingerprints” may vary temporally as new ICD codes accrue for an individual patient over time. Nevertheless, it is critical to ensure that processes are put in place to mitigate this risk e.g. by confidentiality agreements that explicitly prohibit re-identification, currently not required by HIPAA when none of the 18 identifiers are included. This is particularly relevant in the context of the licensing of large clinical data sets in recent years to third party research data brokers.

References
Contextual Inquiry of the Research Data Flow of the Greater Cincinnati/Northern Kentucky Stroke Study

Paul Murdock¹, Brett Harnett, MS¹, Kathleen Alwell, RN², Dawn Kleindorfer, MD², Brett Kissela, MD, MS², Danny T.Y. Wu, PhD, MS¹

¹University of Cincinnati Department of Biomedical Informatics, Cincinnati, OH; ²University of Cincinnati Department of Neurology and Rehabilitation Medicine, Cincinnati, OH

Introduction

The Greater Cincinnati/Northern Kentucky Stroke Study (GCNKSS hereafter)¹ aims to improve the understanding of trends in stroke incidence and mortality. Since 1993, GNCKSS has been monitoring and analyzing temporal trends in the incidence rate, causes, treatment, and outcome of stroke in the Greater Cincinnati area. While this clinical research has made several scientific contributions, the amount of data in this study is voluminous with significant manual effort, resulting in a tremendous amount of time needed to determine case eligibility. In this informatics project, we aim to improve the efficiency of the research data flow of GCNKSS and conduct a qualitative study to understand its work processes and efficiency issues.

Methods

We conducted 15 semi-structured interviews to collect contextual data regarding the research data flow of GCNKSS. The participants were recruited from the GCNKSS research team through a research coordinator. During the interviews, the participants were asked to describe their workflow and act out their everyday project tasks. Workflow issues and bottlenecks were explicitly asked as one of the questions. The interviews were audio recorded and transcribed verbatim by the first author. To construct the research data flow and identify its bottlenecks, the first and the last author reviewed the transcriptions and coded the work processes described and issues mentioned by the participants. The codes were discussed to produce a workflow diagram with bottleneck issues highlighted. In this abstract, we focus on the five key team members (a research coordinator, a data manager, an experienced research nurse, a senior faculty physician, and a lead biostatistician) in the research data flow. This study received an exemption from the Institutional Review Board at the University of Cincinnati.

Results

Figure 1 shows the results based on the interview data from the five key team members. The total length of the five interviews was 240 minutes, with each interview ranging from 30 to 90 minutes. Four bottlenecks include 1) collecting patient list based on diagnostic codes from multiple hospitals in the area, 2) cleaning the patient lists to produce a master list for case determination, 3) manual case reviews by physicians, and 4) database quality check.

Conclusion

In this abstract, we conducted qualitative interviews to understand the GCNKSS research data flow and demonstrated preliminary results. We will continue analyzing all interview data in a second round of coding to gain a more comprehensive picture of the efficiency issue to develop a clinical research informatics solution. Lessons learned from this work have a potential to inform the design of informatics solutions of other large scale longitudinal clinical research studies.

References

Jointly Embedding Biomedical Entities and Text with Weak Supervision

Denis Newman-Griffis, Albert M. Lai, PhD, Eric Fosler-Lussier, PhD
1 The Ohio State University, Columbus, OH, USA; 2 National Institutes of Health, Bethesda, MD, USA; 3 Washington University, St. Louis, MO, USA

Introduction: In biomedical discourse, key concepts are referred to with a variety of natural language expressions, often quite different from one another. For example, “amyotrophic lateral sclerosis” and “Lou Gehrig’s disease” both refer to the same neurological disorder. However, word embedding models, which have recently been applied to a variety of biomedical natural language processing tasks to great effect [1], are unable to capture the fact that these two strings refer to the same disorder. Nonetheless, embedding models offer many compelling advantages, including direct comparability between objects, fast computation, and encoding contextual information about the corpus they were trained on. As a result, recent work has investigated learning vector space models of ontology concepts, instead of textual forms, to better capture the semantics of atomic concepts. However, these methods have relied either on associated structured data (e.g., billing codes) [2] or noisy, automated concept tagging methods that discard textual contexts [3]. We present a novel method for learning embeddings of text and ontology entities using weak supervision. We present representations for more than 3 times the number of concepts embedded in prior work, and we demonstrate that our embeddings achieve competitive performance on several biomedical NLP tasks.

Methods: We present a novel method for learning embedded representations for words, phrases, and biomedical concepts from a large, unannotated text corpus. We use weak supervision from the Unified Medical Language System (UMLS), by first retrieving the set of text phrases used to refer to canonical concepts in the UMLS, and then finding exact matches of these phrases in our target corpus. Each phrase match is treated as a proposed occurrence of its associated concepts, weighted by similarity to a fixed-width context window around the phrase. These contexts are then used, along with randomly-selected negative samples, to train a log-linear embedding model for words, phrases, and concepts together, using our own software implementation. We evaluate our embeddings on two intrinsic tasks: the UMNSRS Similarity and Relatedness datasets, using cosine similarity to measure both similarity and relatedness; and BMASSE, a recent dataset of biomedical analogies, using standard vector arithmetic approaches.

Results: We learn embeddings of over 180 thousand UMLS concepts, using the 2016 PubMed baseline. We find that our concept embeddings outperform both word embedding methods and prior concept embedding methods on the medical similarity (0.65 vs 0.57 Spearman’s ρ) and relatedness (0.56 vs 0.55 Spearman’s ρ) tasks, though we do not match knowledge-based methods that explicitly leverage the UMLS structure. On the biomedical analogy dataset, concept embeddings improve performance on several challenging relations, including gene-encodes-gene-product (36% concept accuracy vs 2% word accuracy), and noun-form-of (27% concept accuracy, 3% word accuracy). In addition, preliminary results on the task of biomedical word sense disambiguation (WSD) indicate that concept embeddings represent a promising direction for this task. Through an analysis of our learned embedding space, we also find that elements of the UMLS ontology structure, such as clustering of semantic types and the similarity of concepts and their text forms, are preserved with some regularity in our embeddings.

Conclusions: Our results indicate that we capture semantic characteristics of ontology concepts as well or better than prior methods that require structured data on top of an unannotated corpus. Furthermore, we are able to scale to a much larger concept vocabulary than prior methods without additional expense. Finally, while our embeddings only partially retain the structured knowledge of the UMLS, our joint learning method means that they both encode information about the text contexts they are mentioned in and are directly comparable to unseen text documents.

This research was supported by the Intramural Research Program of the National Institutes of Health, Clinical Research Center and through an Inter-Agency Agreement with the US Social Security Administration.

References
Justin Norden, MPhil\textsuperscript{1,†}, Christy Tomkins-Lane, PhD\textsuperscript{4,†}, Aman Sinha, MPhil\textsuperscript{3,†}, Richard Hu, MD\textsuperscript{5}; Mathew Smuck, MD\textsuperscript{1,2}

\textsuperscript{1}Stanford University School of Medicine, Stanford, CA; \textsuperscript{2}Department of Orthopaedic Surgery, Stanford University School of Medicine, Stanford, CA; \textsuperscript{3}Department of Electrical Engineering, Stanford University, Stanford, CA; \textsuperscript{4}Department of Health and Physical Education, Mount Royal University, Calgary, Alberta CA; \textsuperscript{5}Department of Surgery, University of Calgary, Calgary, Alberta CA; \textsuperscript{†}Contributed Equally

Purpose: Musculoskeletal disorders are associated with significant mobility limitations. In particular, osteoarthritis (OA) and lumbar spinal stenosis (LSS) are two of the leading causes of disability in the US. Accelerometers have the potential to become inexpensive and noninvasive tools for both diagnosis and management in clinical settings. In particular, accelerometry data could be useful to objectively describe activity in daily-living conditions. However, despite the increased popularity of accelerometers in wearable devices, the clinical relevance of accelerometry data remains unclear. Our objective is to illustrate the explanatory power of accelerometry data in describing performance in clinical populations. Specifically, we present a method that uses accelerometry data to characterize and distinguish between three populations: healthy individuals, individuals with lumbar spinal stenosis (LSS), and individuals with osteoarthritis (OA). Motivated by clinical applications, our data-driven approach rigorously creates an easily interpretable “phenotype” of accelerometry-based features to distinguish (pairwise) between these populations.

Methods: All analyses were conducted using existing datasets, including: the Osteoarthritis Initiative, the National Health and Nutrition Examination Survey (NHANES) 2003-4 data, and the Lumbar Spinal Stenosis Accelerometry Database (LSSAD). In order to characterize the accelerometry signals of OA, LSS and NHANES, we examined the data using 1) standard intervals from Freedson et al., and 2) the Profiles of Physical Performance (PoPP) analysis designed by our group specifically to interrogate data from mobility-limited populations. We then determined the statistical significance of each feature in discriminating between populations, as measured by the p-value of the logistic regression coefficient. Finally, we generated a sparse set of the most relevant discriminatory features by performing regularized logistic regression over the entire set of statistically significant features.

Results: Using the NHANES 2003-4, OAI, and LSSAD datasets, on average 84\% of features were found to be statistically significant in discriminating between each pair of populations at p<0.05. Furthermore, we generated a sparse set of features that discriminate between the groups at 80\% accuracy given age and gender. The most important distinguishing features corresponded to sedentary and light activity. On the other hand, the subtler classification between diseased populations was at 72\%, with moderate activity as the prominent distinguishing feature.

Conclusion: We have illustrated the potential relevance of accelerometry data in distinguishing between disease-specific populations. Specifically, our novel data-driven approach formalizes the correlation between pain and physical activity through accelerometry-based features, and it determines the most discriminatory of these features. This technique demonstrates how to derive clinical relevance from accelerometry data. Our approach determines a key set of discriminatory features, resulting in a framework for classifying musculoskeletal diseases. We have created a quantitative phenotype that provides a more comprehensive analysis of free-living movement.
Factors Associated with Hospital-Acquired Catheter-Associated Urinary Tract Infections Using Multiple Data Sources and Logistic Regression Modeling

Jung In Park, PhD, RN¹, Donna Z. Bliss, PhD, RN, FAAN, FGSA², Chih-Lin Chi, PhD, MBA², Connie W. Delaney, PhD, RN, FAAN, FACMI², Bonnie L. Westra, PhD, RN, FAAN, FACMI²

¹School of Medicine, Stanford University, Stanford, CA
²School of Nursing, University of Minnesota, Minneapolis, MN

Catheter-associated urinary tract infections (CAUTIs) are associated with increased morbidity, mortality, and health care costs. There are a number of clinical guidelines about hospital-acquired CAUTIs (HA-CAUTIs), but the rate of HA-CAUTI occurrence is still rising. There is a need to investigate additional factors that have been associated with the development of HA-CAUTI. The Knowledge discovery and data mining (KDDM) method offers an approach to discover new insights from a large-scale data. This study employed a KDDM approach using logistic regression (LR) to model the dichotomous variable of patients with HA-CAUTIs and patients without HA-CAUTIs.

The purpose of this study was to develop a model using LR to identify factors associated with HA-CAUTI. The setting was three adult intensive care units (ICUs) — medical, surgical, and cardiovascular — in a Midwest university hospital. The study was a secondary data analysis and three datasets were integrated: Electronic health record data extracted from Clinical Data Repositories was combined with staffing and environmental data from the hospital’s National Database of Nursing Quality Indicators and a list of patients with HA-CAUTI. The total number of patients was 8,496, and the total number of unique ICU admissions for analysis was 11,226.

The LR model result showed the effect of a single variable, indicating what factors were associated with a higher risk of HA-CAUTI. Results showed that the factors associated with a higher risk of HA-CAUTI were: young-old and old-old adult, Charlson comorbidity index score ≥ 3, longer length of stay, glucose lab result higher than 200 mg/dl, presence of rationale for continued use of catheter, higher percent of direct care RNs with associate’s degree in nursing, and higher percent of direct care RNs with BSN, MSN, or PhD. The factors associated with lower risk of HA-CAUTI were: young adult, middle-aged adult, male gender, immunosuppression, Charlson index score ≤ 2, hospitalization within previous 6 months, pre-existing urinary catheter, total nursing hours per patient day, and percent of direct care RNs with specialty nursing certification.

The results support the association of the following factors with HA-CAUTIs found from previous studies: female gender, older age (> 50); severe underlying disease; longer length of stay; glucose lab results (> 200 mg/dl); longer use of the catheter; and RN staffing. The results suggest that removing urinary catheters as soon as possible, having more RNs with specialty nursing certification, and developing gender-, and age-specific protocols for catheter care are needed to reduce HA-CAUTI occurrence. Findings also suggest that presence of patients with HA-CAUTI in the unit requires more nursing care from higher level nurses.
Using the PARAFAC2 Tensor Factorization on EHR Audit Data to Model Clinician Behavior

Ioakeim Perros, M.Sc.1*, Xiaowei Yan, Ph.D.2, J.B. Jones, Ph.D.2, Jimeng Sun, Ph.D.1, Walter F. Stewart, Ph.D.2

1Georgia Institute of Technology, Atlanta, GA; 2Research Development & Dissemination, Sutter Health, Walnut Creek, CA; *Work conducted during an internship at Sutter Health

Objective

Electronic health records (EHR) data are increasingly used to understand and improve care processes within and between encounters and to monitor progress towards quality and efficiency targets. While clinic practice process improvement (e.g., Lean principles) is often used in such efforts, the application of these methods in healthcare is far less successful than in other business sectors, in part, because of gaps in scientific process measurement. In fact, a common criticism of healthcare process improvement work is the lack of measurement principles supported by readily accessible continuous data.

The EHR’s automatically generated, time-stamped activity log data tracks all clinical transactions of users and is a useful resource for measurement of healthcare delivery processes. However, activity log data are inordinately complex and difficult to organize into intuitive representations of work tasks. In this work, we describe one method for rapidly translating activity log data into work task information.

Materials and Methods

We analyzed two days of the EHR activity log file data on 876 Primary Care Physicians across 16,613 unique ambulatory care encounters. The EHR activities were represented by 2328 unique activity identifiers and a corresponding activity description. We applied a tensor factorization method called PARAFAC2 to identify and segment unique activity log record sequences as PARAFAC2 can handle high-dimensional, sparse and irregular input data. We organized the audit log file as follows: first, we grouped activities occurring within one second of the previous activity into a single observation; this resulted in 48 unique features. We then represented the input data as a binary multi-way array (tensor) of size \( I_k \times J \times K \), where \( I_k \) corresponds to the number of observations of the \( k \)-th encounter, \( J \) is the number of features (i.e. 48), and \( K \) is the number of encounters. The PARAFAC2 model decomposes the input to three interpretable factors: a matrix representing the clustering of features across all encounters, a matrix containing the importance measure of each clustered feature group for each encounter, and a matrix reflecting the temporal variation of clustered feature groups for each encounter.

Results

Without prior information, PARAFAC2 automatically identified 4 representative clinician target tasks: 1) medications’ access, 2) notes’ access, 3) order entry access and 4) diagnosis modification, along with the most prevalent variation pattern for each target task. We then focused on the notes’ access task and discovered 9 distinct patterns of activity associated with note-related tasks. We also exploited the encounter representations reflecting the usage of each variation pattern and identified two distinct physician user groups based on their notes’ access workflow. The two groups accessed notes by either using the Visit Navigator or the Wrap-Up option, implying varying workflow strategies across those groups.

Discussion

Our approach offers: a) the most representative target tasks of the input encounters, along with their corresponding workflow variation patterns (i.e., various ways leading to the same target task), and b) a succinct representation for each encounter reflecting the usage of each variation pattern. These representations can be used to further group clinicians based on their individual patterns of task completion in the EHR. Our results demonstrate the usefulness of the proposed methodology towards transforming the EHR audit file data into work task information. Our work is one of the first attempts to exploit the wealth of the EHR audit file data in order to improve healthcare processes.
Recruiting for a Pragmatic Trial Using the Electronic Health Record and Patient Portal: Successes and Lessons Learned

Emily R. Pfaff, MS¹, Adam Lee, MBA¹, Robert Bradford, BS¹, Jinhee Pae, BA¹, Clarence Potter, MS¹, Paul Blue, BA¹, Patricia Knoepp, MPH¹, Kristie Thompson, MS², Darren A. Dewalt, MD, MPH³
¹NC TraCS Institute, UNC Chapel Hill, Chapel Hill, NC, USA; ²Cecil G. Sheps Center for Health Services Research, UNC Chapel Hill, Chapel Hill, NC, USA; ³Division of General Medicine and Clinical Epidemiology, UNC Chapel Hill, Chapel Hill, NC, USA

Introduction
Querying electronic health records (EHRs) to find patients meeting a study’s criteria is an efficient method of identifying potential participants. Directly messaging potential participants using the EHR is a newer practice, made possible by the advent of patient portals. We are exploring this recruitment method in the ADAPTABLE study (Aspirin Dosing: A Patient-centric Trial Assessing Benefits and Long-Term Effectiveness), a national pragmatic trial using the PCORnet Clinical Data Research Network (CDRN). We aimed (1) to devise a secure method of tracking a single participant’s identity through several systems and (2) to determine whether patient-portal-driven recruitment is an effective method of study recruitment.

Methods
To identify potential participants, we wrote a computable phenotype for the ADAPTABLE eligibility criteria, resulting in a pool of 22,529 patients. These patients were divided into four recruitment groups: (1) Patients scheduled to come in for an appointment → recruited in clinic; (2) Patients with an active My UNC Chart account → receive a My UNC Chart message; (3) Patients without My UNC Chart accounts who have a known email address → receive an email message; (4) Patients with neither a My UNC Chart account nor an email address → receive letters. Other than those in group 1, most patients also received a follow-up phone call five to seven days after the initial contact.

A key challenge is tracking a participant through multiple systems, so that their EHR data can eventually be merged with their study data. Patient medical record numbers (MRNs) were used as one of the cross-system identifiers, but only after encrypting each MRN using the Triple Data Encryption Algorithm (TDEA). Before recruitment began, each MRN was encrypted and stored in a crosswalk table alongside the unencrypted MRN and other study identifiers.

If a patient clicks the link in the My UNC Chart message, they are taken to a REDCap survey. To determine who is filling out the survey (to connect the initial invitation to study enrollment), we transmit the patient’s identity from My UNC Chart to REDCap using the encrypted MRN to protect patient privacy.

Results
The UNCHCS patients recruited to the ADAPTABLE trial as of September 27, 2017 are tallied in Table 1.

Table 1. UNC ADAPTABLE recruitment through September 27, 2017

<table>
<thead>
<tr>
<th>Method of Contact*</th>
<th># Contacted</th>
<th># Golden Tickets Used¹</th>
<th>% Golden Ticket</th>
<th># Enrolled</th>
<th>% Enrolled</th>
</tr>
</thead>
<tbody>
<tr>
<td>In clinic</td>
<td>376</td>
<td>94</td>
<td>25</td>
<td>69</td>
<td>18</td>
</tr>
<tr>
<td>Letter/Call</td>
<td>409</td>
<td>14</td>
<td>3.4</td>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>My UNC Chart Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Message only</td>
<td>1640</td>
<td>212</td>
<td>13</td>
<td>90</td>
<td>5.5</td>
</tr>
<tr>
<td>Message &amp; phone follow-up</td>
<td>30</td>
<td>1.8</td>
<td>12</td>
<td>0.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>82</td>
<td>11</td>
<td>78</td>
<td>4.8</td>
<td></td>
</tr>
</tbody>
</table>

¹“Golden Tickets Used” indicates that the patient logged into the study website, but they may have later declined to participate.
* Data on the efficacy of the email recruitment pathway are still being collected as of this writing, and should be available shortly.

Discussion
Our experience with ADAPTABLE reveals that electronic recruitment is significantly more effective than sending letters, a traditional low-touch recruitment method. Though electronic recruitment is less effective than high-touch, in-clinic recruitment, the potential resources saved make this workflow well worth incorporating into an overall recruitment strategy.
Evaluating a National Early Warning Score (NEWS) based Best Practice Alert (BPA) on Patient Outcomes

Matthew Phelan, MS¹, Cara O’Brien, MD², Armando Bedoya, MD MMCi²
Benjamin A. Goldstein, PhD¹,³
¹Center for Predictive Medicine, Duke Clinical Research Institute Durham, NC, USA; ²Department of Medicine, Duke University NC, USA;
³Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

Abstract

Clinical deterioration in the hospital is often preceded by abnormal vital signs. In 2012, the Royal College of Physicians derived, validated, and implemented the National Early Warning Score (NEWS) in the United Kingdom which is a composite score based on vital sign measurements commonly found as structured data elements in an Electronic Health Record (EHR). In March 2015, the NEWS and a nursing workflow was implemented into the EHR on all adult patients at our institution. At a predetermined threshold, NEWS would trigger a BPA directed to the patient’s care nurse to take further clinical action. Post-implementation assessment noted little change in clinical outcomes, widespread ignoring of BPAs by care nurses, as well as concern on whether or not our early warning system was beneficial to patient care.

In this poster, we explain the considerations in setting up EHR data to appropriately compare the effectiveness of the NEWS implementation by assessing patient deterioration in an inpatient setting. We define patient deterioration as an unanticipated transfer to an Intensive Care Unit (ICU) or death. Because an inpatient encounter can vary extensively (duration of encounter, surgery, location), comparing event rates where number of inpatient encounters is the denominator would be ineffective. Careful consideration must be taken in defining patient time by selecting locations of interest, for example only locations where an event should be considered meaningful. Assessing the effectiveness of a BPA trigger also contains challenges as 86% of BPAs were ignored.

An additional consideration in comparing rates is the temporal aspect and capturing variations in other elements such as medical practices or population differences. After NEWS implementation, there was no decrease in ICU transfers or death, suggesting no positive impact on patient outcomes, Table 1.

Table 1. Changes in Outcomes after Implementing NEWS

<table>
<thead>
<tr>
<th></th>
<th>ICU Transfers/ Death</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total Events</strong></td>
<td></td>
</tr>
<tr>
<td>Pre-NEWS</td>
<td>1092</td>
</tr>
<tr>
<td>Post-NEWS</td>
<td>1051</td>
</tr>
<tr>
<td><strong>Event Rate (events per 100-patient days)</strong></td>
<td></td>
</tr>
<tr>
<td>Pre-NEWS</td>
<td>0.71</td>
</tr>
<tr>
<td>Post-NEWS</td>
<td>0.67</td>
</tr>
<tr>
<td>Relative Change (95% CI)</td>
<td>-6.7% (-15.0%,1.7%)</td>
</tr>
<tr>
<td>Absolute Change (95% CI)</td>
<td>-0.05 (-0.1,0.1)</td>
</tr>
</tbody>
</table>
Estimating Availability and Completeness of Electronic Health Record Data for Pragmatic Clinical Trials
Laura G. Qualls MS 1, Thomas A. Phillips MA1, Jennifer White PhD 1, Mary Williams 1, Katy Knowlin MHA 1, Holly Robertson PhD 1, Lesley H. Curtis PhD 2,3, William S. Jones MD 1, Bradley G. Hammill DrPH 1,3

Author Affiliations:
1. Duke Clinical Research Institute, Duke University School of Medicine, Durham, North Carolina
2. Department of Medicine, Duke University School of Medicine, Durham, North Carolina
3. Department of Population Health Sciences, Duke University School of Medicine, Durham, North Carolina

Introduction
Pragmatic clinical trials frequently leverage real-world data sources such as electronic health records to complement primary data collection. In the National Patient-Centered Clinical Research Network (PCORnet®), network partners create and maintain DataMarts which include standardized, quality-checked electronic health record (EHR) and other data.1 DataMarts are updated on a regular schedule, but the availability and completeness of data in recent months may vary due to differences in source data availability, data processing time, and other factors. Therefore, we developed methods to define DataMart-specific data censoring dates for use in trial-related statistical analyses.

Methods
Data completeness was determined by comparing the actual volume of encounters in the most recent year to the expected volume of encounters. Expected volume was determined by calculating the average monthly volume during a benchmark period. The benchmark period was defined as the period from 12-23 months prior to the current month (i.e., prior year). Data completeness was reported as a percentage of expected volume. For these calculations, we focused on encounters in ambulatory, inpatient and emergency department settings, which are the most relevant settings for many outcomes. We defined the max date as the most recent month with any of these encounters. We defined the censoring date as the most recent month during which the observed encounter volume was at least 75% of expected volume. An alternative cut-off of 90% was explored. We defined data latency as the difference in months between the month the DataMart was updated and the censoring date. Data latency included two components: availability latency (months in which data were not available) and completeness latency (months in which data were available but incomplete).

Results
In August 2017, we analyzed data completeness among 15 DataMarts which participate in the ADAPTABLE trial.2 Data latency ranged from 1-10 months. Availability latency ranged from 0-10 months, and completeness latency ranged from 0-2 months. Adjusting the completeness criteria to 90% would have had minimal impact on the determined censoring date.

Discussion
Adequately accounting for variability in data availability and completeness is an important step in ensuring that results from real-world data are valid and reliable. We found that completeness latency was less variable than availability latency, suggesting that EHR encounter data is fairly complete within 2 months. Future work will focus on measuring data availability and completeness in other domains such as vitals, medications, and laboratory results.

References
1. www.pcornet.org
Context Aware workflow compared to traditional alerts for Clinical Decision Support for In Visit Patient Screening

Matt Rafalski, MD, John Gill, MD, Luke Barron
Heart of Texas Community Health Center, Waco Texas Waco Family Medicine Residency, Waco Texas

Introduction: Traditional Patient screening in EHRs has been accomplished by presenting the screening staff, be it nursing or providers, with an alert, interruptive or non-interruptive, reminding the staff to perform the task. Clinicians would then have to go to a screening section of the EHR and perform and record the screen, then resume their prior workflow. Traditional CDS using some form of alert or reminder has been shown in outpatient community settings to have inconsistent results (1). A great deal of work seems to go into getting traditional decision support to be successful (2). We have done some preliminary work in this study demonstrating that context sensitive workflow can quickly improve screening rates without the use a traditional pop up alert system.

Design: Our Traditional Depression Screening involved the staff looking to a Heath Maintenance alert section of our EPIC EHR, where an icon would alert them as to the need to perform a depression screen on that patient. They would then open a screening section of the EHR to ask and record the results of the depression screen (PHQ-2). Using a rule that evaluated current patient data combined with a trigger in the EHR navigator section display controls, we set the section that displayed the depression screen to appear base on the patients age, whether they had a diagnosis of depression already (these were excluded), and whether a depression screen had been done in the past year. We collected screening data from patients seen the 1st week of each month seen by 90 providers at 14 of our clinic sites. We used clinical quality measure CMS-2 to determine parameters for our numerator and denominator – both our rule and reporting were based on these specifications.

We compared our screening rates 3 ways – to last year’s average screening rate, to 1st quarter prior to improvement in screening rates, and to screening rates in a department that maintained the old workflow.

Results: A total of 33,960 eligible patients were seen in 2016. An average of 14,484 eligible patients were seen per quarter in 2017. Effective Screening rates (patients screened who had a visit during the measured time period) were almost immediately improved over rates from 1st quarter of the current year (pre-workflow change) and from the previous year in departments where the traditional CDS were switched to context aware CDS with inline workflow. This screening rate actually improved with use rather than staying level over 3 quarters of data, suggesting that the improvement was only partially due to education effect. Supporting this was the lack of improvement in the department that did not switch from context aware CDS. Additionally, the context-sensitive CDS is designed to detect previous year’s screens, so the technique shows promise for both new and follow up screening in the future.

Conclusion: Context Aware clinical decision support with inline workflow showed improvement form baseline over traditional CDS compared to previous years/quarters in departments that switched to context aware CDS, and showed improvement compared to lack of improvement compared to department that did not switch to context aware CDS.

Future Work: Implement additional screens with minimal education; show sustained performance in second and third years of screens; compare methods with a cognitive workload scale.


Rapid Interoperable Implementation of Quality Measurement Across Multiple Sites Using the PCORnet Data Infrastructure

Hanieh Razzaghi, MPH1, Kathleen E. Walsh, MD, MSc2, Levon Utidjian MD, MBI1, Shannon M. Alford, MPH3, Susannah Jones2, Rahul A. Darwar, MPH1, David M. Hartley, PhD, MPH2, Kathryn McAuliffe, MPH1, Keith Marsolo, PhD2, Charles Bailey, MD, PhD1
1The Children’s Hospital of Philadelphia, Philadelphia, PA; 2 Cincinnati Children’s Hospital, Cincinnati, OH; 3 University of Florida, Gainesville, FL

Introduction
The National Quality Forum (NQF) recently endorsed a series of quality measures focused specifically on improving pediatric care. One such measure, *Metabolic Monitoring for Children and Adolescents on Antipsychotics*, assesses recommended screening for side effects of this class of medications. In this study, we evaluate the feasibility of implementing the metabolic monitoring measure across multiple health institutions using the PCORnet data infrastructure.

Methods
The NQF claims-based measure is defined as the percentage of children and adolescents 1-17 years of age who had two or more antipsychotic prescriptions (denominator) and who received appropriate metabolic testing (numerator), as defined by at least one glucose and one cholesterol test during the measurement year. The study team collaboratively developed technical specifications for the metabolic monitoring screening measure based on publicly available NQF documentation, accounting for the structure of the PCORnet CDM and differences in network data provenance and capture, requiring several adaptations to the NQF measure. The PEDSnet Data Coordinating Center (DCC) wrote base queries that retrieved data for potentially eligible persons using the PCORnet SAS standard and distributed these to participating institutions, and developed analytic code in R that evaluated the measure based on the established technical specifications.

Results
The SAS code that identified the cohort for evaluation was initially run against 16 sites in the network and 8 sites (n=7491) were included in the final analysis. One was excluded due to data governance issues and the remaining 7 were excluded due to missing measurement data (n=6) or < 5 evaluable patients (n=1).

We tested several modifications to the NQF measure based on its application to the PCORnet data model and potential limitations of the NQF measure (Figure 1). The developed ‘extended measures’ consider the following modifications: 1) addition of codes identified by the study team that were not included in the NQF value sets; 2) extension of the acceptable window for metabolic monitoring to a 3-month window outside of the evaluation year; 3) flexibility (> 1 days apart) in counting distinct drug events for inclusion into the cohort.

Conclusion
We have successfully adapted and implemented a pediatric quality measure within the PCORnet data network infrastructure that is automated and interoperable. The nuances of the PCORnet data model and the disparate data sources had to be accounted for when specifying the measure. With these adaptations and logical extensions, measure performance was notably better than the base NQF specification. These results are potentially of interest to health system leaders for potential benchmarking or performance reporting.
Electronic Health Record Connected Virtual Hypertension Management

Authors
Sheridan Rea; Alexander Sandhu, MD, MS; Nancy Shin, PharmD; Vanessa Guzman, PharmD; Marcie Levine, MD; Megan Mahoney, MD; N. Lance Downing, MD.

Stanford University School of Medicine, Stanford, CA

Introduction
Hypertension is second only to tobacco use as a leading cause of preventable death\(^1\). With just over 50% of Americans with hypertension at goal pressures, a 10% increase in management would save an estimated 14,000 lives per year\(^2\). The current clinical model, despite being informed by adequate knowledge of effective treatment, falls short of managing hypertension for several reasons\(^3\,4\). First, in addition to possible misdiagnoses due to white coat hypertension, intermittent blood pressure (BP) monitoring in clinic can lead to a major risk of adverse events due to low BP outside of clinic. Second, the in-person practice to establish a diagnosis and achieve target pressures places an undue burden on both the patient and the health care system and significantly contributes to patients lost to follow-up and resultant missed opportunity to achieve target pressures. Home hypertension monitoring and management has proven an effective option to improve clinical outcomes and increase patient and provider experience\(^5\). However, challenges integrating patient-collected data into clinician workflow has limited widespread adoption even with national leaders such as Kaiser Permanente. This project aims to examine the clinical outcomes and usability a virtual management model using home BP cuffs that integrate with patients’ smart phone and electronic medical records. Integrating home data into clinicians’ usual workflow, via the electronic health record (EHR) may lead to increased patient engagement, provider efficiency, and lower resource utilization.

Methods
We are actively recruiting uncontrolled hypertensive patients from Stanford primary care clinics in Santa Clara and Palo Alto. Eligible patients are randomly assigned 1:1 to either usual care, consisting of pharmacist-led hypertension management, or usual care plus the ability to transmit home BP data, automatically, to our EHR. Both groups receive an identical BP cuff. When to take BP measurements will be instructed by the pharmacist as per usual care, and we will not be providing extra reminders outside of this. We will be able to see the time of the measurements in the EHR.

The primary outcome is the change in BP of patients over a 6-month period in the control group vs the intervention group. Secondary outcomes include frequency of patient visits to clinic, patient remote interactions, number of medication changes, and time to reach BP control. We will also be examining patient and provider satisfaction with using a connected device and investigating barriers to use. Stanford University’s IRB determined this study to be quality improvement work.

Expected Results and Future Directions
We anticipate that using a connected BP cuff integrated with the patient’s electronic medical record will be acceptable and preferable to usual care to both patients and providers and potentially allow clinicians to titrate medications faster and with fewer clinic visits. This may lead to greater decreases in BP or faster times to controlled BP. Qualitative analysis of patient and provider satisfaction will likely provide information on barriers to use and how this technology can be better optimized to achieve higher usage rates and greater reductions in BPs.

Possible avenues of future inquiry include algorithm based titration and video conference appointments and targeting special populations for whom close BP monitoring is important, such as post-stroke, peripartum, and chronic kidney disease. This information may be valuable to vendors and health systems seeking to deploy EHR-connected devices in new clinical care delivery models.

References
Towards Discerning Purulent from Non-Purulent Skin and Soft Tissue Infections from Veteran Health Administration Clinical Texts

Jamie W. Rhoads, MD, MS\textsuperscript{1,2}, Chao-Chin Lu, PhD\textsuperscript{1,2}, Danielle L. Mowery, PhD\textsuperscript{1,2}, Vanessa Stevens, PhD\textsuperscript{1,2}, Brian Sauer, PhD\textsuperscript{1,2}, Matt Samore, MD\textsuperscript{1,2}

\textsuperscript{1}IDEAS 2.0 Center, George E. Wahlen Veteran Affairs Medical Center; \textsuperscript{2}School of Medicine, University of Utah; Salt Lake City, UT

Introduction

Guidelines for treatment of skin and soft tissue infections (SSTIs) from the Infectious Diseases Society of America were created to promote best practice. Recent studies have raised doubts about some of the guideline elements, e.g., whether to prescribe adjuvant antibiotics for mild purulent SSTIs and whether streptococci should be the targeted organisms in non-purulent SSTIs. Comparative effectiveness studies can be leveraged to determine best practices; however, such studies require large cohorts of patients. The Veteran’s Health Affairs Corporate Data Warehouse provides access to an adequate population, but the ability of ICD-10 codes to differentiate purulent from non-purulent SSTIs has yet to be determined. Our long-term goal is to develop an automated system that accurately discerns whether a patient has a purulent or non-purulent SSTI using clinical text. Our short-term goals are to 1) determine the logic to support inference of purulence based on SSTI descriptions, and 2) identify exclusionary scenarios missed by ICD-10 billing codes.

Methods

We identified ambulatory encounters between 10/1/15 and 9/30/16 associated with an ICD-10 code for SSTI. To identify incident cases, we excluded cases with a preceding SSTI ICD-10 coded within a 60-day window. To limit our population to SSTIs managed in the outpatient setting, we excluded cases with admission to the hospital in the 24 hours following the SSTI event. We used ICD-10 coding to exclude similar diagnoses outside the scope of this project. We randomly sampled 20 SSTI events with the ICD-10 billing codes: L02 (4; 20%), L03 (15; 75%), L08.0 (0; 0%), L08.89 (0; 0%), and L08.9 (1; 5%). We reviewed 104 clinical notes annotating sentences with mentions of explicit purulence (“palpable fluctuance”), uncertain purulence (“possible abscess”), implicit non-purulence (“no pus”). To limit the annotation task, we did not annotate explicit non-purulent mentions e.g., cellulitis, assuming all SSTIs mentions are non-purulent unless otherwise stated. For evidence of missed exclusionary diagnoses, we annotated exclusionary scenarios as suspected (“history of folliculitis”) and confirmed (“will treat for folliculitis”). We report examples of exclusionary scenario distributions and these contexts.

Results

We observed 19 mentions of explicit purulence (56%), 5 of uncertain purulence (15%), and 10 of implicit non-purulence (29%). We identified a total of 21 suspected and 15 confirmed exclusionary scenarios (see Table 1).

Table 1. Distribution of observed exclusionary scenarios and their contexts in the pilot data.

<table>
<thead>
<tr>
<th>Exclusionary Categories</th>
<th>Example Scenarios</th>
<th>Contexts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>suspected (s)</td>
</tr>
<tr>
<td>trauma</td>
<td>“infected laceration”(s); “74 yo who had “freak accident”(s)</td>
<td>8 (47%)</td>
</tr>
<tr>
<td>hospitalized</td>
<td>“pt said MD recommended admission, but pt declines”(s)</td>
<td>4 (80%)</td>
</tr>
<tr>
<td>animal bites</td>
<td>“he did not feel anything bite him” (s)</td>
<td>4 (100%)</td>
</tr>
<tr>
<td>chronic ulcer</td>
<td>“patient is concerned he is developing a diabetic ulcer”(s)</td>
<td>3 (75%)</td>
</tr>
<tr>
<td>folliculitis</td>
<td>“folliculitis: bactrim, hot compresses”(s); “h/o folliculitis”(s)</td>
<td>3 (75%)</td>
</tr>
<tr>
<td>not incident case</td>
<td>“follow-up of RLE cellulitis occurring in July 2015”(s)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>21 (58%)</td>
</tr>
</tbody>
</table>

Conclusion

Most purulent mentions are explicit and few uncertain. Exclusionary scenarios are more often suspected than confirmed. The most common exclusionary scenario given was trauma. We are actively annotating a larger corpus and developing a natural language processing system to extract and classify these mentions from clinical note.
Automatic inference of phenotypic features for mortality prediction

Victor Rodriguez1,2, Adler Perotte, MD, MA1, Noémie Elhadad, PhD1
1Department of Biomedical Informatics, Columbia University, New York, NY; 2College of Physicians & Surgeons, Columbia University, New York, NY

Introduction
Clinical records are comprised of heterogeneous, high-dimensional data. Topic models have been used successfully to reduce high-dimensional patient records to a relatively small set of compact, meaningful features (1). UPhehnome (2), a model closely related to Latent Dirichlet Allocation (LDA) (3), was developed specifically to model heterogeneous clinical data. The model generates two outputs: 1) a set of phenotypes defined in terms of multiple data types (e.g. notes, medications, labs) contained in the clinical record and 2) a set of phenotypic profiles—patient-specific distributions over inferred phenotypes. Phenotypes capture the clinical characteristics of pathologies described in the clinical record, while phenotypic profiles provide low-dimensional summaries of patients’ clinical histories. In this work, we aim to evaluate the utility of UPhehnome as a method for distilling low-dimensional, clinically meaningful features from clinical records. Our evaluation is task-based; we train the model on data for a cohort of critical care patients and utilize the inferred phenotypic profiles as features for downstream predictive modeling of a major adverse clinical outcome, patient mortality.

Methods

Data: Our cohort is comprised of adult patients in the Medical Information Mart for Intensive Care (MIMIC-III) clinical database. Date of death data available in MIMIC-III is used to identify patients whose death occurred during admission or within 30 days of discharge. Clinical notes, medications, and labs generated prior to discharge or the date of death will be used as inputs to UPhehnome.

Feature Inference: We derive and implement a variational inference (VI) algorithm for UPhehnome permitting inference of phenotypes (Figure 1) and phenotypic profiles. The algorithm minimizes the Kullback-Leibler divergence between a simple, yet flexible proposal distribution and the posterior distribution defined by the model. This minimization is equivalent to maximization of the evidence lower bound (ELBO) over the variational distribution parameters. The model’s optimal parameterization, including the optimal number of phenotypes, will be identified as that which maximizes the ELBO on a held-out validation dataset.

Predictive Modeling: Each phenotypic profile will be used as a feature vector for downstream predictive modeling using regularized logistic regression. Features associated with significant positive or negative coefficients, will be interpreted as phenotypes indicative of high or low risk of mortality respectively. Performance of predictive models will be evaluated by calculating the areas under the Receiver Operating Characteristic and Precision-Recall curves on a held-out test set.

Results
We have carried out inference on MIMIC-III data with our UPhehnome VI implementation to an ELBO convergence tolerance of $10^{-6}$. Upon inspection, inferred phenotypes were found to have high face validity for clinical concepts. Previous work has found that LDA topics capturing clinical concepts related to an adverse outcome are also highly predictive of it (1, 4). Similarly, in our predictive task we anticipate phenotypes associated with high mortality risk to capture the clinical characteristics of severe and/or advanced pathologies.

References
EHR Phenotyping and Data-Driven Suicide Prevention

Steven Rogers, MD, MS-CTR\textsuperscript{1,2}, Richelle deMayo, MD, CM\textsuperscript{1,2}, Kun Chen, PhD\textsuperscript{2,3}, Fei Wang, PhD\textsuperscript{4}, and Robert Aseltine, PhD\textsuperscript{2,3}

\textsuperscript{1}Connecticut Children’s Medical Center, Hartford, CT; \textsuperscript{2}UConn Health, Farmington, CT; \textsuperscript{3}University of Connecticut, Storrs, CT; \textsuperscript{4}Weill Cornell Medicine, New York, NY

Background
Suicide is the 10\textsuperscript{th} leading cause of death in the US\textsuperscript{1}. Twenty-five suicide attempts are made for every completed suicide and patients commonly make a health visit in the week prior to attempting suicide\textsuperscript{2}. The National Strategy for Suicide Prevention’s “Zero Suicide” Initiative regards suicide as a preventable outcome and seeks to address gaps in health care that interfere with the identification and management of at-risk individuals.

Objectives
The study aims to (1) develop improved means of identifying patients at risk of suicide; (2) leverage multiple de-identified data sources, including EHR, Connecticut’s “all payer” claims, and mortality data to identify suicide risk phenotypes and (3) provide real-time phenotype and predicted risk to clinicians to inform decision-making.

Methods
Eight years of data from Connecticut’s Hospital Inpatient Discharge Database and the Office of the Connecticut Medical Examiner (2005-2012) including 15,246 patients hospitalized for a suicide attempt (defined by ICD.9 “E” and other codes) were used to explore suicide risk\textsuperscript{3}. Candidate risk predictors including demographic, diagnostic, procedural, and hospital discharge disposition variables were initially screened via Mantel-Haenszel test. Survival time after suicide attempt was modeled using Cox’s proportional hazards stepwise regression analysis, with the final model adjusted to include main effect variables whose interaction terms had been selected.

Figure 1: Analytic Framework

Preliminary Results and Next Steps
High-risk patients, defined as those with less than 90% chance of 5-year survival, constituted 4.2% patients, experienced a 4.4x relative risk of suicide (90% CI: [2.65, 6.36]) and a 3-year mortality of 15%. A subset of clinical predictor variables including history of procedures on tongue or esophagus, injury mechanism of hanging and transfer to psychiatric hospital offers potential for targeted actionable intervention. We next plan to validate our model using EHR data from 6 diverse health systems in Connecticut via a repeated random splitting procedure that produces separate model training and model testing samples. The out-of-sample approach will be used to construct receiver operated characteristic (ROC) curves and assess accuracy of risk classification by area under the curve (AUC). In addition, statistical tools such as sparse and low-rank decomposition and matrix completion will be used in feature extraction and construction of risk phenotypes. Risk phenotypes will be reviewed by clinical experts to develop guidance for response by treating clinicians.

References
\textsuperscript{3} Patrick AR, Miller M, Barber CW et al. Identification of hospitalizations for intentional self-harm when E-codes are incompletely recorded. Pharmacoepidemiology and Drug Safety, 2010;19:1263e75
Early Experiences Enabling a Machine Learning Pipeline with Clinical and Medicare Shared Savings Program (MSSP) Data

Ursula A. Rogers¹, Shelley A. Rusincovitch, MMCl¹, Erich S. Huang, MD PhD ¹, Ricardo Henao, PhD², Michael Gao², Mary Schilder, MA¹, Eugenie Komives, MD¹

¹Duke Health, Durham, NC; ²Duke University, Durham, NC

Introduction
It is uncommon that both clinical and claims data are available for the same patient population. Duke Connected Care, an Accountable Care Organization (ACO) participating in the Medicare Shared Savings Program (MSSP), has undertaken a machine learning project which takes advantage of such a unique, combined dataset with Electronic Health Record (EHR) data and Centers for Medicare and Medicaid Services (CMS) claims data to create an actionable predictive model for implementation.

Challenges
With the goal of predicting patient admissions, the team’s initial challenge was supporting this high-performance data science project within a typical enterprise IT environment. Table 1 outlines early challenges encountered in both the data itself and the technical architecture designed to enable machine learning, a regularized neural network classifier, on the combined dataset (~390 GB).

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Description</th>
<th>Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACO attribution</td>
<td>Population in CMS files varies over time based on complex inclusion criteria and attribution</td>
<td>Consider only the patients with CMS claims in the past 2 years; identifying activity ensures 90% attributed are scored</td>
</tr>
<tr>
<td>Patient matching</td>
<td>Patients in CMS population need to be linked to Duke’s EHR patients</td>
<td>Utilizing Epic’s reliable Caboodle patient matching algorithms yields &gt; 89% match</td>
</tr>
<tr>
<td>Encounter basis difficult to match</td>
<td>Claims and EHR processes record visits differently (billing vs. clinical care perspectives)</td>
<td>Date matching performed; combined data is stronger by filling in care gaps (EHR has lab data, claims has out-of-network visits)</td>
</tr>
<tr>
<td>Latency</td>
<td>There is generally a 2-month lag time in the effective date of CMS data when received, contrasted with EHR data that is available less than a day old</td>
<td>10 years of the most current EHR data along with 2 years of claims data, recognizing the lag in availability, positively improves model performance</td>
</tr>
<tr>
<td>High volume computing</td>
<td>Large combined datasets, in conjunction with high-throughput modeling, requires significant compute power</td>
<td>Potent computational servers are utilized for machine learning, with secure access to the enterprise environment (see Figure 1)</td>
</tr>
</tbody>
</table>

Table 1. Challenges in working with high volume EHR and claims datasets.

Innovation
Figure 1 illustrates the technical design for operationalizing the admission prediction model. The MSSP datamart resides within the enterprise database so that it can access raw data and serve up extracted/matched datasets for model consumption. In turn, the model writes output risk scores to result tables that are available for various types of user reporting. Access is tightly locked down via honest broker due to identified data.

Conclusion
The combination of the claims and EHR datasets are proving to be a powerful resource in predicting patient admissions and informing patient care.

Figure 1. Implementing a machine learning model within an enterprise workflow.
**Common Data Models: When Playing with FHIR, You Don’t Have to Choose Just One**  
Leon Rozenblit, JD, PhD, Patrick Young, PhD, Philip Langthorne,  
Prometheus Research, LLC, New Haven, CT

**Introduction**

Many initiatives have formed to promote the creation and adoption of common data models (CDMs) for medical informatics, driven by a need to integrate data from numerous sources to support benchmarking, effectiveness research, and evidence-based practice improvements. Example initiatives with active CDM specifications include the National Patient-Centered Clinical Research Network (PCORnet), Informatics for Integrating Biology and the Bedside (i2b2), the Sentinel Initiative, and the Observational Medical Outcomes Partnership (OMOP). Organizing data into any of these CDMs offers transferability of data within the CDM network, but it does not solve the problem of transferring data to other models that may be more advantageous to use at different times for their own reasons. For example, although the Sentinel and PCORnet CDMs provide limited content coverage versus OMOP, PCORnet in particular counts dozens of partner organizations and the PCOR Institute additionally offers research funding, which is not true of other CDM initiatives. Meanwhile, using a limited-coverage CDM as one’s primary information model restricts one’s ability to repurpose the data to richer CDMs, but using an overly-inclusive CDM risks over-complicating simple inquiries and may not align with researchers’ own mental models. Ideally, a clinical data repository will be successful in supporting transferability of data among both sources and uses of the data.

**Method**

We approached the problem by collecting data from electronic health record systems using the emergent transport standard from Health Level Seven, Fast Healthcare Interoperability Resources (FHIR), which we selected equally for its rich semantic model as for its robust syntactic rules. The FHIR resources are aggregated in a PostgreSQL database exactly as they were received, as JavaScript Object Notation (JSON) objects, using an entity-attribute-value schema with limited, frequently-traversed references (such as “Patient” and “Encounter”) modeled as unary relationships on the same table, for processing convenience. We then translated the data into the PCORnet CDM through a Python-based mapping process, and presented it in a queryable data mart for addressing clinical research questions using standard data analysis tools such as SQL, SAS, SPSS, Excel, and other such programs. Separately, we also translated the FHIR data into an in-memory representation adhering to the Quality Data Model (QDM), in order to support simultaneous calculation of clinical quality measures for benchmark reporting back to care providers and for submission to the Center for Medicare and Medicaid Services. In both cases, the original FHIR resources were preserved in the central data warehouse and remained available for additional translations.

**Results and Discussion**

Using FHIR as our information model for both transport and persistence allowed us to aggregate granular, semantically rich clinical data into a central repository, where each data object remained faithful to its exact source encoding but adhered to a common structural model. Separating the most convenient CDM for analysis of the data (PCORnet vs. QDM) from the information model used for storage allowed us to support multiple simultaneous applications (clinical research vs. quality improvement) requiring different representations of the data that would be difficult to reconcile without the common FHIR representation. This multi-model approach provided a significant advantage over the problem of “CDM lock-in” and may prove useful in other clinical data applications as well.

**References**

2. https://www.i2b2.org/
3. https://www.sentinelinitiative.org/
Using Clinical Quality Improvement and Clinical Informatics Principles to Decrease the Medication Management Build Time

Connie L. Saltsman, Pharm.D., MBA, CPHIMS, Risa C. Rahm, Pharm.D., CPHIMS
HCA Healthcare, Nashville, TN

Introduction

One of the larger sections of a multi-disciplinary software implementation for hospitals is the Medication Management build. Medication management is extremely complex, and if medications are not built and implemented correctly, they could cause major harm to patients. With all of the improvements in Electronic Health Records and new software availability, clinical informaticists are constantly evaluating new options, and are a part of the evaluation, design, build and implementation of new systems. Being able to decrease the cost, time and complexity of the Medication Management build is important to be able to install these new applications appropriately.

Project

The Clinical Pharmacy Informatics Team identified that a more streamlined method was needed in order to consistently and efficiently implement new clinical applications across the 170+ facility enterprise. We executed a step wise approach to address this problem, using Clinical Quality Improvement (CQI) techniques, as well as Clinical Informatics (CI) Principles in each of the steps. The Plan-Do-Study-Act (PDSA) model was used, as well as Root Cause Analysis (RCA) and Failure Modes and Effects Analysis (FMEA) in different portions of the project. We also utilized data governance and stewardship, human factors engineering principles, data standardization, and knowledge application and delivery principles throughout the entire project.

The first step was to identify standards for over 20 Medication Management dictionaries since our EHR build was not standard across the enterprise. The second step was to work with the pharmacy informaticists across the company and identify medication management build best practices. Step three utilized other projects and triggers to update the build and roll out those best practice standards.

The fourth step was to utilize feedback from providers, pharmacists, nurses, other clinicians, and informaticists to evaluate the impact of the previous steps. This provided us with enough data to acquire additional funding for further standardization and reduction of medication build ambiguity. This included two projects focused on the pediatric/neonatal population and the medication reconciliation process. Another project that was funded included provider-facing Medication Management Clinical Decision Support (CDS). As part of this, we have been able to build and deploy over 100 standard CDS.

The latest step in the process was to identify what product enhancement change requests could be made in the clinical application to better streamline the Medication Management build. A multi-disciplinary team, including the vendor, reviewed the build and implementation project plan step-by-step to determine what was working well, what could be standardized in the implementation approach, and what could be done quicker with software enhancements. We are now working with the vendor to put together requirements and a timeline for these enhancements.

Conclusion

The use of CQI and CI principles greatly enhanced our ability to decrease the Medication Management build and implementation time during this step-wise approach. We have utilized multi-disciplinary design sessions to evaluate each step within the project. Each step was evaluated to determine if it included as much standardization, lessons learned and feedback from clinicians as possible.

In our current step, working closely with the vendor, we were able to collaborate to understand the development needs and timelines for our ideas, and could choose the enhancements that optimized the value both for the time and cost invested. Our next steps are to continue to work with the vendor to build requirements documents for the enhancements and prioritize the product development.

We will show in the poster examples of the various CQI and CI principles used in our projects, as well as data that shows the decrease in build time from the initial EHR implementation where this process was begun, and how we have utilized the approach in our current EHR implementation to identify opportunities to decrease the build and validation timeline.
Designing and Implementing Safe IT Practices as Part of Clinical Workforce Education
Elizabeth L. Schoff
School of Population Health, University of Auckland, Auckland, NEW ZEALAND

The purpose of this project was to reduce overall cybersecurity risk in a large regional health board by helping staff develop good digital hygiene. A critical risk identified was that of introducing malware (malicious software) from emails with hyperlinks known as phishing.\(^1\) The project aimed to reduce this risk by quantifying staff digital habits and then improving these metrics through education campaigns, including online training.

Health systems are particularly at risk for being compromised by cyber criminals who exploit vulnerabilities in information technology (IT) for monetary gain. These risks exist for many reasons. The technology in the health sector is often complex and software development practices have not emphasised security. Poor security practices in the diverse user base have also contributed to these risks. Time-constrained staff use hospital email accounts for non-hospital activities, viewing websites and emails that may contain links to malware.

Health IT is now frequently used for successful health interventions. Clinicians depend on these systems. Safe, trustworthy and accessible information systems are now mission critical. And yet, the most vulnerable aspect in keeping information systems secure is the behaviour of the very people who depend on these systems to deliver information to the right place, at the right time and to the right person.\(^2\)

**Methods:**
This new initiative followed the ISO Security Standard 27001:2013 three step approach of (1) define a policy, (2) educate staff, (3) evaluate. We used a co-design pedagogical approach.\(^3\)

We used 3 evaluation steps. (1) Baseline measurement via Phriendly Phishing using an email campaign to entice staff to click links that could have been associated with malware. (2) Communication campaign in hospital newsletter articles, screen savers, and emails, plus an invitation to participate in online training module. (3) Repeat step 1 in Feb 2018.

**Results:**
Nearly 20% of the 26,000 staff clicked links that could have compromised the network during the baseline campaign. Some staff contacted the IT Service Centre with concerns about emails that included links, but said they were not sure what action to take.

The uptake for the training, which was optional due to staff time-constraints, was 47.5%. Some senior clinical staff felt their knowledge in other areas (medicine) made them cyber-aware and that training was unnecessary. The final analysis for risk reduction will be conducted following the campaign ending in February.

The original policy was the typical IT centric ‘User Acceptance Policy.’ However, a re-branding to ‘Safe IT Practices’ made it more relevant for staff as they were already familiar with the necessity of having ‘Safe Clinical Practices’. Connecting secure networks with patient safety and good data hygiene maintained the clinical focus.

**Conclusions:**
Clinicians generally do not see themselves as ‘computer people’, and therefore the growing issue of cyber-crime is often viewed as an IT Department problem. Terms like ‘cyber-security’ do not resonate with clinical staff, making it more difficult to create awareness for these risks.

Despite public reports of incidents like ransomware compromising hospitals, it is challenging to create a culture of safe IT practices. Making IT security education compulsory and including it within CME (Continuing Medical Education) may help integrate this new skill set needed for effective clinical practice.

\(^1\) “Sixth annual Benchmark study on Privacy & Security of Healthcare Data”; Publication Date: May 2016; downloaded from: https://www.ponemon.org

\(^2\) Goedert, Joseph, “Providers’ vulnerability to email-borne malware is still high” Publication Date: December 12, 2017; downloaded from: https://www.healthdatamanagement.com/

A Scalable Approach to Support Multiple Cohort Discovery Projects in i2b2

Evan T. Sholle, M.S.1, Marcos A. Davila, M.S.1, Joseph Kabariti, B.S.1, Julian Z. Schwartz, B.S.1, Thomas R. Campion, Jr., Ph.D.1,2

1Information Technologies & Services Department, Weill Cornell Medicine, New York, NY; 2Department of Healthcare Policy & Research, Weill Cornell Medicine, New York, NY

Introduction
The secondary use of patient electronic health record (EHR) data for cohort discovery is increasingly common in academic medical centers. One popular approach is i2b2, which stores data from EHR and other systems in a star schema and allows investigators to run queries in a user-friendly web-based tool. After creating one i2b2 instance containing de-identified data for 2.7 million patients at Weill Cornell Medicine (WCM), we deployed i2b2 to meet the needs of specific groups of investigators requiring identified data for patient subsets of interest to them. However, we soon encountered disk space and extract-transform-load (ETL) time constraints. In response, we implemented a scalable approach based on structured query logic (SQL) views and compared it to the standard table-based approach.

Methods
After setting up individual i2b2 projects for separate investigator groups, we extracted source system data, transformed it into the i2b2 star schema format, and loaded it into a single base i2b2 project; assigned specific cohorts of patients to each project; created SQL views to support each project; and created indexes optimized for SQL views in i2b2. Metadata objects specific to individual projects remained as tables in the view-based approach, but patient data objects shared between projects became views. We then compared the standard table-based approach to the novel view-based approach with regard to disk space and ETL time required. We used the Microsoft SQL Server 2014 database.

Results and Discussion
As described in Table 2, we estimate the standard table-based approach required upwards of 10 terabytes of disk space and over 57 hours to complete for a base project containing all patients and five projects containing specific cohorts of patients. Because patients who existed in cohort-specific projects also existed in the base project, the table-based approach created identical patient records and facts across multiple databases. Almost half of the disk space and ETL time required for each project was attributable to index creation.

Table 2. Space and time required to conduct ETL using standard vs. new approach.

<table>
<thead>
<tr>
<th>Project</th>
<th>Number of Patient Records Created (Millions)</th>
<th>Number of Patient Facts Created (Millions)</th>
<th>Disk Space Required under Standard Approach (Terabytes)</th>
<th>ETL Time Required under Standard Approach (Hours)</th>
<th>Disk Space Required under New Approach (Terabytes)</th>
<th>ETL Time Required under New Approach (Hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base (all patients)</td>
<td>2.78</td>
<td>1,725</td>
<td>4</td>
<td>22</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>Digestive care</td>
<td>0.25</td>
<td>760</td>
<td>1.76</td>
<td>9.68</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Anesthesiology</td>
<td>0.24</td>
<td>806</td>
<td>1.86</td>
<td>10.27</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.03</td>
<td>214</td>
<td>0.49</td>
<td>2.73</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Myeloproliferative neoplasms</td>
<td>0.07</td>
<td>409</td>
<td>0.94</td>
<td>5.21</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Ophthalmology</td>
<td>0.19</td>
<td>586</td>
<td>1.35</td>
<td>7.48</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>3.50</td>
<td>4,503</td>
<td>10.44</td>
<td>57.41</td>
<td>4.05</td>
<td>27</td>
</tr>
</tbody>
</table>

Compared to the standard approach, we estimate the view-based approach reduced required disk space by 61% and ETL time by 53%. The approach has enabled us to scale multiple i2b2 projects across our institution and may generalize to other settings. Future work can extend the approach to other database platforms such as MySQL. This study received support from the National Center for Advancing Translational Sciences (NCATS) Clinical and Translational Science Center (CTSC) (UL1 TR0002384) and WCM Joint Clinical Trials Office (JCTO).
Supporting Queries of Microbiology Results in i2b2 with Culture, Organism, and Antibiotic Sensitivity Data

Evan T. Sholle, MS¹, Joseph Kabariti, BS¹, Shari S. Gelber, MD, PhD², Thomas R. Campion, Jr., PhD¹,³,⁴

¹Information Technologies and Services Department, Weill Cornell Medicine, New York, NY; ²Department of Obstetrics and Gynecology, Weill Cornell Medicine, New York, NY; ³Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY; ⁴Department of Pediatrics, Weill Cornell Medicine, New York, NY

Introduction

Clinical and translational researchers often seek to query microbiology results with respect to culture type, organism, and sensitivity¹. However, optimal approaches for representing microbiology result data in an easily queryable form are unknown. Responding to investigator demand, we developed a novel approach for representing microbiology data in i2b2, a popular platform that enables secondary use of electronic health record (EHR) data for patient cohort discovery. In contrast to existing approaches, we sought to develop a hierarchy that allowed for queries that returned all cultures that identified a particular organism, regardless of their source.

Methods

We created two i2b2 ontologies for culture and organism. The culture hierarchy displayed specific microbiology orders from our institution, such as “BLOOD CULTURE” or “PREGNANCY PANEL,” with modifiers for each order representing the full range of organisms that each culture has identified. In contrast, the organism hierarchy displayed names of organisms, with modifiers for each organism representing the cultures in which the organism has been identified. The organism hierarchy also displayed antibiotic sensitivity modifiers. Each hierarchy was generated using existing data from Weill Cornell’s EHR, showing only results that have returned for a specific ontology item.

Results

Initial pilot sessions with faculty indicated that the ontology is intuitive and well-suited to addressing investigator use cases with scientific value. Compared to existing approaches¹, our method has enabled users to view all potential organisms that have been identified from a particular culture, as well as all cultures that have ever identified a given organism without having to run a query. In conjunction with i2b2’s “Selected groups occur in same financial encounter” option, users can restrict their query to only return patients who tested positive for organisms with resistance or susceptibility to a particular organism. Users can thus run queries as specific as “How many patients tested positive in 2016 for vancomycin-resistant E. faecium in a blood culture?” In contrast to previous approaches, which focused on orders and offered organisms as they returned from particular culture types, our approach allows users to query by order and organism with or without respect to the other. For example, using Chandaka et al.’s ontology, a user could not query how many patients tested positive for E. faecium regardless of source – in order to do so, they would have to identify each order that had ever returned results positive for the organism, a prohibitively difficult task. Under our ontology, they could simply locate the organism in its hierarchy, then drag it in to query all patients who tested positive for it, regardless of the order associated with the result.

Other institutions may find the approach useful for representing microbiology results in i2b2 to support clinical and translational researchers. Future work will include formal evaluation of the ontology’s soundness, completeness, and usability. Additionally, inclusion of a standardized ontology for organisms, such as SNOMED, may facilitate future investigator queries, as well as federation of data with information from other institutions. This study received support from NewYork-Presbyterian Hospital (NYPH) and Weill Cornell Medical College (WCMC), including the Clinical and Translational Science Center (CTSC) (UL1 TR000457) and Joint Clinical Trials Office (JCTO).

References

**Title:** A Case for Data Mining: Predictive Analytics Using Medicare Part D Data  
**Author:** Jaime Smith, MAE  
**Affiliation:** Department of Health Administration and Policy, George Mason University – PhD Student

**Introduction:** The use of prescription opioids by elderly people in the U.S. has grown considerably, mostly driven by increased rates in prescription drug addiction, efforts to address chronic pain in the elderly, and industry initiatives that promote the use of prescription drugs (Anupam, 2014). Additional research has identified other factors as likely contributors to the severity of the current prescription drug abuse problem, including: drastic increases in the number of prescriptions written and dispensed, greater social acceptability for using medications for different purposes, and aggressive marketing by pharmaceutical companies (Paulozzi, Strickler, Kreiner, & Koris, 2015). However, until the recent focus on the opioid abuse epidemic, much of the attention on opioid abuse has centered on patient behavior with little to no emphasis on prescribing patterns of providers.

By 2014, there were no national estimates on the frequency of opioid prescribing by health care providers for elderly populations (Anupam et al., 2014); however in the following year, CMS released its first-ever Medicare Part D PUF in an effort to increase transparency and availability of Part D prescription fill data to the public. Research in this area has focused on opioid prescriptions for the general population, without a specific focus on Medicare Part D beneficiaries (Anupam et al., 2014). Despite this gap, there are limited studies using Medicare Part D data and even fewer using data mining techniques and classifiers to predict prescribing patterns of Medicare providers who write prescriptions for controlled and non-controlled substances for patients under the Medicare Part D Prescription Drug Program (Anupam et al., 2014).

Therefore, this study examines the prescribing behavior of U.S. Medicare providers who prescribe opioids to Medicare Part D beneficiaries; it investigates the relationships between Medicare Part D prescriber demographics and their overall prescribing of opioids over the period of 2013-2014; and it uses data mining techniques and descriptive analyses to develop and validate classification models that predict the change in opioid prescribing for Medicare Part D providers between 2013 and 2014. Additionally, this analysis considers the relationship between whether or not an eligible provider has attested for the Meaningful Use EHR Incentive Program prior to and including the 2014 attestation year, and prescribing behavior of Part D providers. Meaningful Use is defined as using certified EHR technology to improve quality, safety, efficiency, and reduce health disparities; engage patients and family; improve care coordination, and population health; and maintain privacy and security of patient health information (CMS HealthIT.gov, 2015).

**Methods:** Descriptive statistics using 100% of the prescriber-level Medicare Part D public use files were calculated to describe the following: 1) Overall Medicare Part D prescribing trends in 2013 and 2014; 2) Prescribing of opioid prescription fills; and 3) Medicare Part D beneficiary counts, total drug costs, provider specialty, and provider state. This analysis partitioned the sampled data into a train/test set where 70% of the sample was designated for training the predictive model while the remaining 30% was used to test the model. Three separate machine learning techniques popularized for use with discretized data were performed: The J48 decision tree classifier, Naïve Bayes, and Logistic Regression. All three methods were trained and the results were reported and compared using the Area Under Receiver-Operator Curve (or AUC) (Huy, et al., 2013). Each technique has its advantages, the J48, decision tree is a predictive machine-learning model that creates a binary tree based on various attribute values of the available data; for logistic regression, a linear classifier is used for supervised learning which has properties such as feature selection and robustness to noise; and the Naïve Bayes classifier works on a simple, but comparatively intuitive concept and often outperforms many other complex algorithms (Deshmukh & Patil, 2011), (Patil & Sherekar, 2013). The AUC is the probability “that a randomly chosen positive instance in the test data is ranked above a randomly chosen negative instance, based on the ranking produced by the classifier” (I. H. Witten, 2011). Here, the AUC is the probability of correctly classifying increases (or non-increases) in prescriptions between 2013 and 2014 based upon the relationship between the true and false positive rates for each classifier. The ideal ranges for the AUC were 70%-100% and the closer the AUC is 1, the better the model (Westra, et al., 2011). Additional analyses will be conducted using SAS version 9.4. Statistical significance was measured using the threshold of P < 0.05.
Results: As shown in Figure 3, more than 50% of all prescriptions filled by Pain Management, Interventional Pain Management, Hand Surgery, and Anesthesiology Medicare Part D specialists were opioids. At least 15% of total prescriptions for many surgical specialties (such as Orthopedic Surgery, Neurosurgery, Maxillofacial Surgery, Oral Surgery (dentists only), Oral & Maxillofacial Surgery, and others) are for opioid prescriptions. In comparison, although Family Practice and Internal Medicine specialties account for the majority of filled opioid Part D prescriptions, only 5.3% of total claims from Family Practice specialists are opioids, and only 4.3% of the total claims for Internal Medicine.

Geographically, the states of CA, NY, and FL led the way with the greatest number of Part D prescription fill claims in the US. Additionally, CA, FL, and TX beneficiaries filled the greatest number of opioid claims over the study period. Figure 4 illustrates the percent opioid claims by provider geography. OK, NV, and AL had greatest prescriber-level opioid rates by state (between 7.3% and 7.4%, respectively). NY, DC, and HI were among the lowest for opioid rates (2.5%-3.2%).
In order to construct the predictive models for this analysis, a 5% random sample of the final dataset was created in SQL Server. This included the 2013 and 2014 Medicare Part D Prescriber data files and the Meaningful Use Attestation data for prescribers who attested prior to (and including) year 2014. As a result of sampling, more than 57,000 Part D providers were used to develop train and test datasets for this analysis. Data-mining methods (J48, Naïve Bayes, and Logistic Regression) were computed using a 70/30 split, train and validation strategy. When predicting 2014 opioid_increase, each of the three models selected resulted in correctly classified instances ranging from 75.9% to 76.8%. (Table 3). The weighted average true positive rates ranged from 0.759 to 0.768 with false positive rates ranging from 0.321 to 0.328. The AUC for the J48 Decision Tree Classifier was 0.744, followed by the AUC for the Naïve Bayes method at 0.782. The logistic regression model, however, demonstrated the highest AUC, at 0.821. Overall, provider specialty and volume of non-opioid prescribing were the top features in the model, Meaningful Use attestation was not significant.

Conclusions: The findings from this analysis identified multiple data mining techniques, such as J48, Naïve Bayes and Logistic Regression that deliver accuracy and results that may be used to enhance clinical decision making for healthcare decision makers. As the opioid epidemic has remained at the forefront of public awareness, public health agencies and policymakers are working behind the scenes to develop strategies to curb overprescribing, treat patients who are already addicted, predict which patients will likely become addicted, and analyze and predict the prescribing behavior of opioid prescribing providers. The public health implications may suggest that targeted programs toward Medicare providers may provide additional support to Medicare patients and the prescribers in informing them on the potential best practices for taking and prescribing prescription opioids. For policymakers, our findings may suggest that additional monitoring of identified specialists may contribute to reducing fragmented care across Medicare Part D recipients. Future studies may leverage this research to explore state-level analyses of Medicare provider prescribing patterns. Finally, as data become available through prescription drug monitoring programs (PDMP) and through federal or state incentives, researchers will have broader access to more robust data for analysis.
TITLE: Initial experience with an informatics-based population management program for men with bladder cancer at the Greater Los Angeles Veterans Affairs Medical Center

AUTHORS: Nicholas Smith MD, Christopher Saigal MD, MPH, Gloria Wong NP, Stanley Ho MS, Jeremy Shelton MD, MSHS

(a) David Geffen School of Medicine at the University of California Los Angeles
(b) Department of Urology, Health Services Research Group, David Geffen School of Medicine at UCLA, Los Angeles, California
(c) Division of Urology, West Lost Angeles VA Medical Center, VA Greater Los Angeles Healthcare System

INTRODUCTION AND OBJECTIVE:
Effective care for non-muscle invasive bladder cancer exists and has been demonstrated to prevent disease recurrence and progression, however compliance is often poor. We leveraged existing information technology resources at the Greater Los Angeles VA Healthcare System to create a clinical dashboard to identify and track all patients followed by the division of urology with bladder cancer in near real-time as an aid to ensuring receipt of guideline-recommended care through an informatics-based disease management approach. Clinical management was performed by a nurse practitioner with physician oversight. Here, we describe our initial experience.

METHODS:
To identify patients with bladder cancer we used an iterative process to develop an algorithm for care deficiency detection using a combination of administrative and clinical data elements including International Classification of Diseases Ninth Revision (ICD-9), Current Procedural Terminology (CPT) codes, VHA facility and clinic codes, and VHA death and facility attribution tables. We used the dashboard log books and the VA data warehouse to examine bladder cancer care between 9/2012 and 9/2016, the total number and percentage of patients contacted, their responses, and the outcome.

RESULTS:
Between 9/2012 and 9/2016, 832 patients were managed by the division of urology. During this time, 211 patients (25%), and 186 (22%) unique patients, were flagged by the clinical dashboard as missing one or more aspects of guideline recommended care. 73 (8.8%) unique patients were confirmed on chart review to be missing care and contact was attempted, while 65 already had scheduled follow-up, 31 had transferred care, 17 had died in the 3 years prior to starting the program, 1 died during the study period, and 3 had disease that did not need follow-up. Of the 73 patients who were contacted after being flagged for non-compliance, 52 (6.2%) were successfully rescheduled for and seen in clinic, 10 agreed to be seen in clinic but repeatedly no-showed, 4 declined care, 4 died, and 3 transferred care. Of the 52 patients recalled to clinic, 23 received intra-vesical immunotherapy, 16 received surveillance cystoscopy, 2 received repeat transurethral resection, and 11 other forms of management.

CONCLUSIONS:
Our intervention findings suggest that an informatics-based population management approach for bladder cancer patients improves compliance with evidence-based, guideline recommended care. Importantly, many of the 52 patients detected and recalled received critically important interventions upon return to clinic. This will presumably reduce recurrence and progression of disease, as these interventions have been shown to do so elsewhere. Our study, however, highlights the challenge of inaccurate or incomplete data and the complexity of case management. Further work is in process to evaluate the full impact of this informatics tool on our population from the perspective recurrence and disease progression.
Identification of Quality Measures in Diabetes Mellitus Patients for MACRA Using Natural Language Processing

Ergin Soysal, M.D., Ph.D.\textsuperscript{1}, Anupama Gururaj, Ph.D.\textsuperscript{1}, Hua Xu, Ph.D.\textsuperscript{1}
\textsuperscript{1} School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX

Abstract

Clinical Quality Measures are playing more important role in U.S. healthcare. Recently, MACRA has tied payment models to clinical outcomes and resource use to promote high-quality patient care. On the other hand, collection of quality data is a major obstacle since this requires prominent changes in clinical processes. We aimed to identify and extract quality measures data for Diabetes Mellitus from free text narratives in combination to structured EHR data and achieved an overall F1 score of 0.90

Introduction

Recently, The Medicare Access and CHIP Reauthorization Act (MACRA) puts a focus on healthcare quality more than ever before. Several quality measures have been proposed and implemented to monitor and measure care quality in clinical practice. These regulations in U.S. healthcare system make quality data collection is a crucial step in healthcare process. On the other hand, considerable amount of these quality data is embedded in EHR text narratives, and requires manual processing. Collection of these data becomes an important problem in chronic diseases with long term follow-up. Natural language processing (NLP) offers appropriate approaches to gather these data. We aimed to identify clinical quality measures using Diabetes mellitus (DM) as a use case.

Methods

We selected a group of 100 DM patients, and split them into development and test subgroups of 50 each. All the clinical encounter/visit, consultation and progress notes for each patient extracted and categorized based on note types. The development subgroup was used to develop a pipeline using CLAMP (Clinical Language Annotation, Modelling and Processing Toolkit) to detect existence of dilated eye examination for retinopathy, examination of feet for diabetic foot and peripheral neuropathy, and any evaluation for diabetic nephropathy. A keyword list was generated to detect examination types in relevant sections of clinical notes. Regardless of existence or inexistence, any mention or reference to a related problem was accepted as positive for that examination type. The test group manually reviewed and annotated to use for evaluation of NLP pipelines for the existence of particular examination type.

Results

Narratives from test group are processed using the developed pipeline, with focus to identification of particular examination type. A total of 1679 notes from 50 patients were processed to locate examination types for target systems to be affected from DM, and results are summarized in table 1. Diabetic retinopathy results were slightly lower than expected, since ophthalmology department notes were missing and it only relies on mentions in the other notes.

Table 1. Results of identification of examination type, with precision, recall and F1 values.

<table>
<thead>
<tr>
<th>Examination Type</th>
<th>n</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diabetic Foot</td>
<td>50</td>
<td>0.95</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>Peripheral neuropathy</td>
<td>50</td>
<td>0.97</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td>Diabetic retinopathy</td>
<td>50</td>
<td>0.85</td>
<td>0.67</td>
<td>0.75</td>
</tr>
<tr>
<td>Diabetic nephropathy</td>
<td>50</td>
<td>0.89</td>
<td>0.92</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Conclusion

Collection of quality data from clinical narratives is an expensive and time consuming process. Natural language processing (NLP) promises an inexpensive, fast and reliable way of extraction of relevant clinical measure data. Especially, in chronic diseases with long term follow-up such as diabetes mellitus patients This study shows that, NLP can take responsibility in collection of clinical quality measures.
Predicting Compliance with Laboratory Utilization Alerts by Binary Logistic Regression and Random Forest

Jeffrey Szymanski, MD, PhD, Abraham Qavi, MD, PhD, Ronald Jackups, MD, PhD

Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO

Introduction

Alerts for unnecessary laboratory testing in the computerized provider order entry system can be an effective method to reduce wasteful orders, but compliance with these utilization alerts is often poor or unpredictable. For any specific alert, the provider's response may be influenced by numerous characteristics of the patient, provider, or targeted test. Identifying the factors influencing compliance can improve alert design. However, laboratory order alerts are typically studied for only one or several tests and over a short intervention period, making it difficult to draw any broader conclusions across different tests.

Since September 2012, Barnes Jewish Hospital has instituted minimum retesting interval alerts for 20 tests, generating over 21,000 alerts. All alerts appeared in the same format in hospital order entry system and fired whenever repeat testing was ordered in the same patient visit. We used the depth and diversity of this data set to discover robust correlates to alert compliance using binary logistic regression and random forests. The objectives of this study were to optimize overall prediction of alert compliance, and to identify specific features most predictive of compliance.

Methods

Alert and ordering data were collected for 52 months from September 2012 to January 2017. Provider compliance was modeled by binary logistic regression and random forests. Model features were either directly extracted from the laboratory information and hospital ordering systems or engineered based on associated hospital information system data. Features directly extracted included patient and provider demographics (patient age and gender, provider type, etc.), previous test results, and test normalcy. Engineered features including measures of alert and order frequency (such as total previous alerts) and interval (such as time from last alert). Continuous data was either rescaled as fraction of maximum (e.g. for time intervals) or group to categories (e.g. grouping continuous results to normal/low/high).

For random forest, 54 features were used to generate 300 trees with 6 variables per split. For binary logistic regression, all features were considered, and were incorporated into the model via least absolute shrinkage and selection operator. Statistical analysis was performed in R v3.4.1. Random forests were generated with the randomForest package v.4.6, and results were visualized with ggplot2 v.2.2.1.

Results

Over the course of 52 months 21,185 alerts fired, affecting 2-10% of orders depending on test. These alerts covered 3,375 providers and 9,966 patient visits with an overall compliance rate of 46% (SD 11%). For random forests, the alert dataset was divided randomly 3:1 for testing and validation, and for logistic regression, 10-fold cross validation was used.

Most predictive variables from the random forest analysis included time from previous alert, time from previous order, provider class, and previous test normalcy. Most predictive variables from logistic regression also included time from previous order, provider class, and previous test normalcy, and additionally included overall provider alert burden and per-test provider alert burden. Overall model performance was compared by area under ROC curve. Random forest (AUC = 0.78) and logistic regression (AUC = 0.74) performed similarly.

Conclusion

Laboratory utilization alert compliance may be improved by studying features that correlate with past compliance. Assessing these features using random forest or binary logistic regression successfully produced an overall classifier predictive of alert acceptance or bypass. There was substantial overlap in predictive features between the two methods. Incorporating knowledge gained from these predictive features into alert design may improve compliance in the future.
A comparison of machine learning methods for automated EHR phenotyping of stroke patients

Phyllis M. Thangaraj, MPhil, Joseph D. Romano, MPhil, Fernanda Polubriaginof, MD MA, Nicholas Giangreco, BS, Mitchell Elkind, MD, Nicholas P. Tatonetti, PhD
Columbia University, New York, NY, USA

Introduction

Stroke is the second leading cause of death in the world.1 Accurate identification of the disease is essential for risk stratification and time-sensitive treatment but can be difficult and time consuming. Automated identification of stroke patients in the EHR can improve the triage and treatment process. Previous studies have applied machine learning methods to identify various phenotypes, but few have systematically evaluated multiple models with various case and control combinations for each phenotype and for stroke.2,3 Given the inconsistency of diagnosis code implementation and the difficulty of identifying true controls in the EHR, we compare and evaluate the accuracy of forty-five classifiers in distinguishing stroke patients from other forms of cerebrovascular disease.

Methods

We extracted data from longitudinal health records within the Columbia University Medical Center Clinical Data Warehouse (CDW). We gathered all diagnostic and procedure ICD-9 and 10 codes, prescriptions, race/ethnicity, age, and gender of each patient as approximately 50,000 binary features. We trained logistic regression, random forest, and adaboost classifiers on the features gathered for fifteen combinations of cases and controls. Codes used to curate the cases and controls were excluded from the feature set. The case sets, ranging in size from 4484-36378 patients, included: 1) a gold standard sample of patients with acute stroke curated by Columbia neurologists, 2) patients with codes for acute ischemic stroke, and 3) patients with codes for cerebrovascular disease. Controls, 7987-5.3 million patients, included: 1) patients without cerebrovascular disease codes, 2) patients without codes for stroke, 3) patients with codes for cerebrovascular disease other than ischemic stroke, 4) neurology patients with codes for stroke mimetic diseases, and 5) a random sample of patients. Controls were randomly sampled to create a 1:2 ratio of cases to controls and reduce the class imbalance. We implemented L1 regularization and evaluated logistic regression and adaboost models using 10-fold cross validation and random forest models with out of bag scoring. Evaluators were scored by the area under the receiver operating characteristic curve (AUC-ROC) and accuracy. To evaluate classification between stroke and other cerebrovascular disease, we measured precision and recall for classifying a test set with gold standard, stroke diagnosis code, and other cerebrovascular disease patients.

Results

For distinguishing stroke cases from controls, AUC-ROC across models ranged from 0.751 (±.0179) -0.997 (±.00140) and accuracy 0.620 (±.0123)-0.950 (±.0116). For distinguishing stroke from other cerebrovascular disease within the test set, maximum F-score ranged from 0.68-0.75. Models with cases from the gold standard had the highest AUC-ROC and accuracy, regardless of control. Models with controls of cerebrovascular diseases other than stroke had the highest max F-scores, suggesting that they distinguish between stroke and cerebrovascular disease the best. We also showed that manually curated cases classified with similar precision and recall levels as stroke cases identified through diagnosis code only. These results suggest that the phenotypic question drives the most effective model, and manually curated training sets are not always necessary.

Discussion

We have developed and compared several different classifiers to distinguish patients with acute ischemic stroke from other cerebrovascular events. This tool is valuable for identifying patients for research studies. We have shown that structured data is sufficiently accurate for classification, leading to widespread usability of the algorithm. Future directions include determining the smallest possible size of training and feature sets, including unstructured data of all patient medical notes, comparing deep learning methods, and testing reproducibility in an independent data set.

References

Automating Installation of the Informatics from Bench to Bedside (i2b2) Platform

Kavishwar B. Wagholikar, MBBS PhD 1,2,3, Michael Mendis 3, Pralav Dessai, MS4, Javier Sanz, MS, 4 Sindy Law 5, Micheal Gilson, MS, 5 Stephan Sanders, MBBS PhD, 5 Mahesh Vangala, MS, 6 Shawn N. Murphy, MD PhD 1,2,3
1Massachusetts General Hospital, Boston, MA; 2Harvard Medical School, Boston, MA; 3Partners Healthcare, Boston, MA; 4University of California Los Angeles, CA; 5University of California, San Francisco; 6UMass Medical School, Worcester, MA

Abstract
Informatics from the Bench to Bedside (i2b2) is an open source clinical data analytics platform that is used at over 150 institutions for querying patient data. We have created an automated installation package, called i2b2-quickstart, which automatically downloads the latest i2b2 source code and dependencies, and compiles and configures the i2b2 cells to create a working i2b2 hive installation.

Introduction
Informatics from the Bench to Bedside (i2b2) is an open source clinical data analytics platform that is used at over 150 healthcare institutions for querying patient data. The platform comprises of several i2b2 cells that provide different services, and the cells communicate with each other using XML web services. However as the platform has several components, it takes several weeks for new users to read the documentation and create a working installation of the platform. The effort needed to standup a new i2b2 hive installation is a major obstacle for wider utilization of the platform.

Methods
We developed a quick-start package to automatically download and compile the latest i2b2 source code, and to install and configure the components in the Linux environment. We isolated external dependencies that may be unavailable or unsupported in the near future, and hosted them in a cloud environment.

Results and Discussion
We have created the i2b2-quickstart package to automatically download, compile and configure the i2b2 platform to run a functioning demonstration instance. In contrast to the manual installation of i2b2, our quick start package requires a single input parameter, i.e. the Internet protocol address of the i2b2 host machine. We hope that our efforts will significantly reduce the time and effort needed to install the i2b2 platform.

Acknowledgements
This work was supported by National Library of Medicine grant R00-LM011575, and Weill grant at UCSF. Amazon provided free credits for use of their cloud service for this project.
Data Profiling in Support of Entity Resolution of Multi-institutional EHR Data

Pei Wang, PhD¹, Daniel L. Pullen, PhD², Maryam Y. Garza, MMCi¹,
Anita C. Walden, MS¹, Meredith Nahm Zozus, PhD¹

¹University of Arkansas for Medical Sciences College of Medicine, Department of Biomedical Informatics, Little Rock, Arkansas, USA; ²University of Arkansas at Little Rock, Department of Information Science, Little Rock, Arkansas, USA

Abstract

Information quality (IQ) is a multifaceted discipline with many degrees of complexity in implementation, especially in healthcare. Data profiling is one of the simpler tasks that an organization can perform to understand and monitor the intrinsic quality of its data. We discuss the benefits of using data profiling to better understand quality issues and their impact on entity resolution and how data profiling might be augmented to increase utility to clinical data.

Methods

We tested the application of simple data profiling techniques to inform the selection of match fields and inform data standardization performed prior to the record linkage operations of ER. We focused our efforts on: ER between multiple datasets to support secondary use of healthcare data in research. Two unique cases were identified for data profiling and Entity Resolution: (1) data from a community registry to be linked to data from the EHRs of a group of local, federally qualified clinics, and (2) data from the data warehouse of an academic medical center. Data elements considered for use as potential match fields from the registry, EHR, and warehouse data were profiled with SQL scripts written against data in a MySQL relational database. Within the registry data, there were 10,069 records with 30 potential match field attributes. There was a total of 25,924 records and 10 potential match field attributes in the EHR data. There were 1,462,888 records and 20 potential match field attributes in the warehouse data. The data profiling results were reviewed to (1) select match fields, and (2) inform necessary data cleaning and standardization.

Conclusion

Data profiling is a technique used in IQ efforts to better understand the basic quality of data by allowing organizations to analyze intrinsic aspects of data quality at the data element level. We leveraged data profiling techniques to assess candidate match fields between Entity Resolution of a registry and a EHR dataset. We also use the same approach to find candidate match fields within health system warehouse data. The summary statistics generated offer insight into key features of the data, i.e., the format, pattern frequencies, and the number and percentage of null, distinct, and unique values. The data profiling results (measures on intrinsic aspects of data quality) adequately described the aspects of the data impactful to the subsequent ER processes. While the selection of fields for record linkage may seem intuitive, data profiling is still advantageous (and highly recommended) for improving one’s understanding of the data prior to performing the ER. Data profiling allows for the discovery of patterns, inconsistencies, and anomalies in the data that would need to be addressed prior to matching, which may reduce the number of false negatives. For example, in the use cases presented here, the profiling of the date or phone fields identified inconsistencies in data format and type. These discoveries facilitated the decision making process for match rule development. The profiling results supported the selection of the match fields for the ER work and highlighted data elements requiring data cleaning or standardization prior to or during the matching process. Further, the data profiling results were used to inform the data cleaning and standardization algorithms used.
Generating a table of aggregated patient data in the i2b2 web client

Nich Wattanasin, MS¹, Taowi D. Wang, PhD¹, Bhaswati Ghosh, MS¹, Reeta Metta¹, Jay Tarantino¹, Vivian S. Gainer, MS¹, Shawn N. Murphy, MD, PhD¹,²
¹Partners HealthCare, Boston, MA; ²Massachusetts General Hospital, Boston, MA

Introduction

At Partners Healthcare, we have developed a new i2b2 web client plugin to help facilitate both the reviewing and downloading of patient data in an aggregated format. This approach allows for an institution operating a local i2b2 data warehouse to install a plugin within their web clients that enables an investigator to be able to retrieve aggregate data, one-patient-per-line, as specified. The user interface of the plugin allows the investigator to first select a patient cohort to generate a table of data on, and then add additional phenotypic parameters and aggregation options that will represent columns within the table. Such aggregations include displaying the total count of observations, dates of first or last occurrence, average or most frequent laboratory values, etc. Due to the potential, long-running process of both a large number of patients or many columns of aggregate data, the plugin is designed to be able to continue generating the table of data in the background, allowing for the investigator to close their web browsers and return at a later time to retrieve their results.

![Figure 1. The Patient Set Viewer i2b2 web client plugin.](image)

Architecture

The plugin (Figure 1) allows an investigator to carefully define how the table of data will look and what data it will contain. Standard i2b2 constraints such as date and value constraints are also supported in the plugin. When the user clicks the ‘Review Patients’ button, the table specification in JSON format is sent from the web browser to a server-side PHP backend that creates a “job file” for the request, and spawns a “worker process” on a separate thread to start generating the table. The web browser will continue to poll to obtain the status of the job and can be closed at any time. When the user signs-in next time, the polling will continue and the generated table will be displayed if the job has been completed.

Conclusion

This new i2b2 web client plugin approach is the basis of many use-cases and applications, such as reviewing patients for clinical trials. The plugin helps to close the loop on the research of a large number of patients in an i2b2 data warehouse to be filtered down—via both phenotypic and genomic queries in i2b2—to a manageable set of patients for review, directly in the i2b2 web client.
**Mining Medication Administration Counts for AKI Prediction and Feature Discovery**

Samuel J Weisenthal, MS¹, Martin Zand, MD, PhD¹  
¹ University of Rochester Medical Center, Rochester, NY, USA

**Introduction**

Acute kidney injury (AKI), a temporary or permanent decline in kidney function, is associated with increased morbidity (e.g., long-term dialysis), mortality, length of stay, and hospital cost. There is recent work in electronic health record (EHR)-based models for AKI and interest in using machine learning methods for predictor discovery. We complement existing studies by using lasso and gradient boosting classifier (GBC) on count vectors of all past medication administrations in a large cohort of rehospitalized patients. We reveal novel predictors for this cohort and condition.

**Methods and Results**

This study was approved with exempt status by our local review board. AKI was detected in codes and in lab creatinine according to Kidney Disease: Improving Global Outcomes (KDIGO) guidelines. We focused specifically on rehospitalizations (n=124,518). We extracted features corresponding to the count of medication administrations over time. We use iterated, grouped, nested cross validation (CV) with lasso and GBC. Our dataset contained 34,505 patients who generated 90,013 samples with 5,618 cases and 84,395 controls. GBC was selected in each inner fold as the best classifier with ROC AUC Micro: 0.82635 ± 0.00693; Brier Micro: 0.05161 ± 0.00189; and PR AUC Micro: 0.27079 ± 0.01484. Results with lasso were ROC AUC Micro 0.80671 ± 0.00764, Brier Micro: 0.0564 ± 0.0022, and PR AUC Micro: 0.22051 ± 0.01397. We show medications with top coefficients for lasso in Table 1.

**Table 1.** Top positive and negative coefficients from lasso. Each feature corresponds to the count of administrations of medication over a patient’s previous hospitalizations. Mean and STD are micro-averaged over 50 iterations of 5-fold CV.

<table>
<thead>
<tr>
<th>Medication</th>
<th>LR(+) Mean</th>
<th>LR(-) Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUROSEMIDE</td>
<td>8.21</td>
<td>-13.81</td>
</tr>
<tr>
<td>HEPARIN SODIUM</td>
<td>6.75</td>
<td>-7.87</td>
</tr>
<tr>
<td>ALLOPURINOL</td>
<td>4.74</td>
<td>-4.58</td>
</tr>
<tr>
<td>ENOXAPARIN SODIUM</td>
<td>4.43</td>
<td>-3.94</td>
</tr>
<tr>
<td>PIPERACILLIN-TAZOBACTAM IN D</td>
<td>4.05</td>
<td>-3.55</td>
</tr>
<tr>
<td>DEXTROSE</td>
<td>3.94</td>
<td>-2.89</td>
</tr>
<tr>
<td>TACROLIMUS</td>
<td>3.62</td>
<td>-2.34</td>
</tr>
<tr>
<td>METOPROLOL TARTRATE</td>
<td>3.54</td>
<td>-1.96</td>
</tr>
<tr>
<td>HYDRAZINE HCL</td>
<td>3.33</td>
<td>-1.58</td>
</tr>
<tr>
<td>TORSEMIDE</td>
<td>3.05</td>
<td>-1.57</td>
</tr>
</tbody>
</table>

**Discussion and Conclusion**

We reveal novel predictors for AKI. Interestingly, ibuprofen is not predictive because of its explicit effect on AKI risk (Ibuprofen is actually known to physiologically increase AKI risk—by interfering with blood flow to the kidney, but here we see that it has the highest coefficient), but likely because of the patient population to which it is commonly administered—patients with low AKI risk (it is avoided in patients with vulnerable kidneys). Increasing AKI risk are diuretics, allopurinol (given for gout—an excess of uric acid that should be secreted by the kidney), piperacillin-tazobactam, likely for sepsis, which can cause AKI, and hypertension (HTN) medications metoprolol and hydralazine. Decreasing AKI risk are medications related to hospitalization for reasons other than chronic disease: antiemetic ondansetron, prenatal medications, and tetanus-diphtheria-pertussis vaccine for pregnancy, antihistamine promethazine and hydroxyzine, bronchodilator albuterol for asthma attacks, nicotine for smoking cessation, and muscle-relaxant cyclobenzaprine—all associated with temporary visits in patients with robust health. Our study suggests that EHR-based predictive models for AKI might best depend on features—such as ibuprofen administrations—that might not have been chosen a priori for an AKI prediction model. Better predictions could lead to timelier preventative interventions. With replication, this approach might also spur further investigation of certain medications as potentially modifiable risk factors for AKI.
References


A Neural Network Model for the Post-Injury Diagnosis of Mild Traumatic Brain Injury using the Pediatric Quality of Life Inventory

Kevin A. Wilson, MS MPH1,2
1RTI International, Research Triangle Park, NC
2Rutgers Biomedical and Health Sciences, Newark, NJ

Introduction
Pediatric traumatic brain injury (TBI) is a major contributor to developmental issues in children and adolescents, with an incidence of between 250 and 799 per 100,000 children per year1. Given the limited proportion of children that seek medical attention for TBI, particularly mild TBI, it is likely that a significant proportion of mild TBI cases go undiagnosed. It is likely that these children may suffer reduced quality of life from 3 months’ post injury. This research will determine, using a neural network, whether it is possible to diagnose a mild TBI at three months’ post injury using the Pediatric Quality of Life Inventory (PedsQL).2 PedsQL comprises 23 questions in 4 domains: physical, emotional, social and school functioning. Neural networks (NNs) have been shown to effectively model non-linear, complex relationships and they have been shown to significantly outperform logistic regression in discrimination tasks.3

Methods
Data were preprocessed to create an analysis dataset that was used to train and evaluate the NN. Subjects with moderate/severe TBI were removed to leave 437 cases of mild TBI and 133 controls. Oversampling with replacement was used to balance the dataset and 144 incomplete observations were removed, resulting in a final dataset comprising 730 observations. All 23 questions in the PedsQL instrument were used as input features. The NN was configured with 1 input layer, 3 hidden layers, and 1 output layer, and trained using 10-fold cross-validation. Performance was assessed using sensitivity and specificity.

Results and Discussion
There were 296 true positives, 332 true negatives, 77 false positives, and 25 false negatives. Sensitivity was 0.79 and specificity was 0.93, both of which indicate excellent performance. The results suggest that the NN is more effective at identifying negative cases correctly than identifying positive cases correctly, which means that the model is optimized towards minimizing false negatives. Clinically it is more important not to miss any TBI cases, so we would prefer a higher false positive rate and lower false negative rate. High specificity suggests that a negative finding may be sufficient to rule out a TBI. In this case, the pediatrician may focus on other potential causes. If there was a positive finding, this would lend evidence for further investigation, such as the ordering of further tests. A key advantage is the brevity of the instrument – it can be completed directly in a clinical decision support system by the child, parent or doctor, and a potential diagnosis can be obtained immediately.

Conclusions
This study aimed to build a neural network model that could be integrated into a clinical decision support tool to diagnose the incidence of mild TBI based on symptoms assessed by the PedsQL instrument. The results showed that the neural network is an effective modeling approach with high sensitivity and specificity, and suggest that a clinical decision support tool based on the model could be informative in clinical practice. While performance was good, specificity was significantly higher than sensitivity, which may result in a bias towards false negatives. Further research is necessary to assess the impact of sample size and validate the model in the clinical setting.

References
Abstract

This research aimed to determine if the neural network technique can be used to diagnose mild traumatic brain injury (mTBI) in children and adolescents three months after the injury has occurred based on a short quality of life questionnaire. Our results indicated that a mTBI injury could be identified with high sensitivity and specificity, suggesting that the neural network approach may be a useful technique to support clinical decision making in the post-injury diagnosis of TBI.
Prediction of Emergency Department Visits within 30 Days: Modeling Risk with Very Big Data from Electronic Health Records

Yifan Xu, PhD1, Michael W. Dusenberry, BS2, Wei Yao, MS1, Amanda Yoho, PhD1
1IBM Watson Health, Cleveland, OH, USA; 2IBM Spark Technology Center, San Francisco, CA, USA

Methodology
We trained two machine learning models using the IBM Explorys Therapeutic Dataset to predict emergency department (ED) visits within 30 days of any face-to-face hospital encounters. 15.6 million patients were eligible for the study, with 2638 features extracted per patient. The models were trained with Apache Spark MLlib and Apache SystemML over a 249-node cluster. A logistic regression model and a random forest model were chosen for their simplicity, interpretability, and robustness.

The study window was from Jan. 1, 2014 to Dec. 31, 2016. For each patient who had at least one 30-day ED visits within the window, one such visit was randomly chosen, and the last discharge date before the ED visit was marked as the indexed date. For others, a random discharged date within the study window was chosen as the indexed date per patient.

Demographic, healthcare utilization, visit month, diagnoses (285 CCS diagnosis groups), and drug prescriptions (885 level-4 ATC groups) on the indexed dates were cleaned and recorded, resulting in 2638 features plus a binary label per patient. The full data was a 15.6 million by 2639 matrix, which was separated into training/validation/testing set. The training set was down sampled to reduce class imbalance. Furthermore, features that were not significantly associated with the outcome were filtered out from the training set.

Descriptive analysis was done to summarize sample characteristics. A random forest model was trained and its hyperparameters were tuned. A logistic regression model was also trained as a baseline. Model performances were measured with sensitivity and specificity for a range of probability thresholds, as well as the receiver operating characteristic (ROC) curves. Variable importance according to the final random forest model was also calculated.

Results
Medicaid and Medicare patients were more likely to have 30-day ED visits following discharge from face-to-face hospital encounters, especially among African American patients. Female patients between the ages of 20 and 40 had significantly higher rates than male counterparts, while children under 10 and under Medicare had the highest 30-day ED visit rates among all groups.

The study population consisted of 1.79M (11.30%) patients with at least one 30-day ED visits in the study window. The training set consisted of 2.1M patients, of which 50% are positive samples. The validation and testing sets contained approximately 3.1M patients each, with roughly 11.30% positive samples in each set. Among the 2638 predictors, 2062 were strongly correlated to the outcome (adjusted p-value < 0.05). The other 576 predictors were excluded from the training set.

The best performing random forest model had 500 trees (tree depth=25, number of candidate features=50, subsampling rate=0.1). AUC values for the random forest and logistic regression models were 0.842 and 0.833, respectively. The top two predictors, “Number of ED visits last year” and “Number of ED visits in the past”, had significantly higher importance scores than other predictors. Trauma and injury diagnoses were high predictors of ED readmission, despite limiting effects of unexpected ED visits by selecting only features associated with the indexed face-to-face visit.

Conclusions
Overall, we have shown the promise of a large-scale, medical machine learning approach for identifying possible sources of ED overuse that could aid the clinical provider in providing increased quality of care at lower costs, as well as laid the framework for future studies that could benefit from a similar approach. Future work includes: validation of our findings with prospective studies; expanding the feature space, such as adding procedure information; exploring increased model complexity to avoid underfitting; and integrating our results into clinical practice in a manner that directly assists clinical providers in providing improved quality of care.
Systematic Detection of Height, Weight Outliers in Electronic Health Record

Wei Yao, MS¹, John P. Borsi, BS¹, Yifan Xu, PhD¹, Nnenna Ibeanusi, MSPH/MSIS¹
¹IBM Watson Health, Cleveland, OH, USA

Introduction

A growing evidence have indicated that Electronic Health Record (EHR) errors lead to data loss, incorrect entry, display or transmission. Given the wide adoption of EHR in primary care and secondary use, the integrity, reliability, and accuracy of health information is critical to delivering evidence-based care to individuals, creating new knowledge and insights through research and to improving medical outcomes for populations. Often EHR errors are observed on fundamental data elements such as height/weight (H/W) anthropometrics, causing both biologically implausible and patient-level conflicting data. Therefore, there is an urgent demand to detect and reconcile these data errors.

Method

A random sample of 16,808 patients with longitudinal health records was selected from IBM Explorys Therapeutic Dataset. Nonparametric regression was utilized to fit smooth curves of age-dependent H/W growth trends, on sex-specific population data and on individual patient records. It uses a Gaussian moving window, which weights all the local points into the fitted curve, thus preventing being biased by relying on individual data points as references for outlier detection. The model estimates the fitted value and the standard error from the whole curve rather than from one data point and emphasizes on how well the estimated curve fit the true underlying population, which is measured by Confidence Intervals (CI). CI establishes the boundary between the true H/W data points and outliers, and thus can systematically identify all outliers.

Results & Conclusion

On the left subplot (Figure 1), population-wise CI was chosen to encompass the normal H/W ranges in the green band according to national survey data and within the yellow band established by median absolute deviation. The subplot on the right represents each patient with individual-level CI in red and population-level CI in blue.

Our outlier detection framework has proved to identify data entry errors, false auto-population of default values and related EHR system errors, and therefore will further improve the data quality for our users.

Figure 1. Population-level and patient-level outlier detection for height measurement in males

References

Treatment Regimen Discovery and Visualization of the Acute Myocardial Infarction cohort with Diabetes type II in the MIMIC-III dataset

Sean C. Yu, MS, Albert M. Lai, PhD
Institute for Informatics, Washington University in St. Louis, MO, USA

Introduction
The Agency for Healthcare Research and Quality enumerates three categories of quality measures: structural, process, and outcome. While there are structural quality measures which assess providers’ capacity for care, and outcome measures that assess the impact of care, most clinical quality measures are process measures that evaluate whether a particular action was taken for a specific cohort of patients, and thus “reflect generally accepted recommendations for clinical practice”. These process measures, however, do not consider the full care pathway nor relate processes to outcomes where implemented. Unlike clinical workflow analysis, which attempts to understand how users perform tasks (often involving technology), process mining is a burgeoning data science subdiscipline that focuses on extracting knowledge from timestamped event data logs to inform process improvement. Of the several process discovery techniques that exist, Inductive Mining is a recursive, graph-based, divide-and-conquer algorithm that is “currently one of the leading process discovery approaches due to its flexibility, formal guarantees and scalability” [1]. Despite this, a comprehensive literature review of process mining in healthcare found only a single study that employed inductive mining within the healthcare domain [2]. The authors of the literature review also remark on the notable absence of good visualization and visual analytics tools for healthcare process models.

Methods
Patient medication data was extracted from MIMIC-III, an openly available dataset developed by MIT Lab for Computational Physiology comprised of deidentified health data associated with ~40,000 critical care patients [3]. The cohort was chosen to be akin to the denominator population for current and past acute myocardial infarction measures (e.g. ICD9 code 410.21), then further narrowed down by diabetes diagnosis (ICD9 code 250.00). 30-day mortality was used as the outcome measure. Patients who expired during hospital stay were excluded from the population. Inductive mining algorithm was implemented in Python and extended to handle infrequent paths and non-atomic events. Visualization was built using Graphviz and made interactive using Jupyter widgets. Width of paths between events is proportional to the number of patients, and the hue of the path is indicative of the outcome of patients who underwent that part of the process (proportion of patients who died within 30 days of discharge).

Results
Using the described initial conditions, medication data for 232 patients with were extracted from the MIMIC-III dataset. A local-edge culling ratio of 0.05 was used to yield a full medication process tree with 99.6% fitness/conformance rate (95.3% naïve fitness rate). Figure 1 is a small subsection of the full medication process tree which shows that 74 patients did not receive Magnesium Sulfate and 8.11% of those patients died within 30 days of discharge whereas 158 patients did receive Magnesium Sulfate and 3.16% of those patients died within 30 days of discharge ($\chi^2 p=0.0989$).

Conclusions
The highly dynamic and complex nature of healthcare processes poses an interesting challenge for those looking to improve such processes. Process mining and visualization tools allow for data-driven identification of high-level procedural structures, and relating such structures to outcomes assist users in diagnosing problematic processes and in reinforcing favorable pathways. The authors aim to explore underlying reasoning for splits in care pathways, and continue to improve the tool in terms of usability, extensibility, and portability.

References
Reference Mining for Improving Cohort Establishment Method Consistency

Yiqing Zhao, MS¹, Yanshan Wang, PhD², Feichen Shen, PhD², Benjamin Yan, Shulan Tian², Jerry Yang, BS³, Nan Wang, BS³Hongfang Liu, PhD².*

¹University of Wisconsin-Milwaukee, Milwaukee, WI; ²Mayo Clinic, Rochester, MN; ³University of Minnesota-Twin Cities, MN;

Background

Mayo Clinic researchers have used Rochester Epidemiology Project (REP), a medical record linkage system, to establish several famous cohorts. However, currently there is no way to keep track of the cohort establishment methods (including cohort definition, confirmation and attribute extraction) used in cohort studies. This may lead to data discrepancies and research inconsistency in cohort studies. Manual systematic reviews of cohort studies have been done to ensure research consistency but automatic reference mining can serve as a more efficient method to examine past discrepancies and ensure future research consistency.

Methods

Our work proposed a framework that combined reference mining and semantic filtering to identify previous work cited in the reference that studied the same cohort. Reference mining is performed using a web crawler package ‘jsoup’ to parse REP publications published in 2016 that are also available on PMC websites. We first identified sections that mentioned cohort establishment methods, e.g. cohort definition, confirmation and attribute extraction. After retrieving the PMIDs and titles of all references cited in the target sections, we applied a UMLS semantic tagger to identify disease related entities in the titles of the original paper and retrieved reference papers. Similarities of entities contained both in the original paper and retrieved reference papers were calculated and used as a filtering measure before finally retrieve a set of references relevant to the cohort establishment methods mentioned in the original paper.

Results and Evaluation

There were 63 REP 2016 publications available on PMC websites at the time of experiment (June, 2017). Evaluation was conducted using manual literature review. Four college level reviewers was instructed to identify references that are relevant to the cohort establishment methods mentioned in the original paper. Automatic extraction results were compared with gold standard from manual literature review results. Inter-rater agreement is 92.5%. The evaluation results show that our framework could achieve a precision of 96.3%, recall of 85.7% and F-score of 0.88. In addition, a user interactive interface was built using D3.js library to visualize the reference tree.

Conclusions

The proposed framework enables retrieval of references related to cohort establishment methods. This framework can be used to facilitate further extraction of historic cohort establishment methods and create a standardized cohort establishment methods for future references. Therefore this framework will be beneficial to the research societies by promoting consistent methods in cohort definition, confirmation and attribute extraction.
Detecting Mental Stressors from Clinical Narratives Using Natural Language Processing: A Feasibility Study

Vivienne J Zhu, MD, MS,1 Chanita Hughes Halbert, PhD,2 Melanie Jefferson, PhD,2 Richard Wolfe, MS,1 Leslie A Lenert, MD, MS1

1Biomedical Informatics Center and 2Hollings Cancer Center at Medical University of South Carolina, Charleston, South Carolina

Abstract: This study tested the feasibility of a natural language processing (NLP) approach for detecting mental stressors from clinical notes. A total of 197,346 clinical notes for 3,138 eligible prostate cancer patients were used to iteratively develop a lexicon of mental stressors, customized negations, and NLP algorithms. With negations, 1,917 (61.1%) patients were detected having stressor mention(s). Mental stressors are commonly documented in clinical narratives for patients with prostate cancer and could be detected by NLP.

Background: The Institute of Medicine (IOM) recently defined 12 measures of social determinants (SD) to be included in electronic medical records (EMR). However, SD information are usually collected from patients by specific survey-based studies. Thus, SD information has not been routinely measured and is usually documented as unstructured data in clinical practice. A potential approach to obtain SD information could be to automatically extract SD from clinical narratives using natural language process (NLP). This study tested the feasibility of NLP extraction of SD from clinical notes, using mental stressor (one key SD) detection as an example.

Method: We used data from Research Data Warehouse of Medical University of South Carolina (MUSC). Patients who are 18 years or older and are diagnosed with prostate cancer were eligible for this study. Source data include selected clinical narratives for the eligible patients recorded between 01-01-2014 and 05-31-2017, such as the patient’s chief complaint, progress notes, consultant notes, outpatient visit notes, and discharge summary. We used commercial NLP software (Linguamatics I2E version 4.4, Cambridge, United Kingdom) licensed to MUSC to index, parse, and query each research criterion. We developed a lexicon and a set of NLP queries to capture semantic and syntactic representations of mental stressors. A set of confirmed search keywords of mental stressors was provided by domain experts. These initial keywords include stress, anxiety, depressed, not restful, feeling irritable, worry, fear, nervous, hopeless, sad, mood swings/elevation, feeling irritable, GAD (general anxiety disorder), PTSD (post-traumatic stress disorder), impatient, tired, and sleep disturbance. We also generated a list of terms for “false” mentions of stress that relate to stress echocardiogram, stress fracture, stress ulcer of stomach, genuine stress incontinence, and other factors. These terms were used as negations to exclude the “false” mentions of mental stress. For each term, we utilized the I2E morphologic variants functions and I2E built-in ontology to generate a set of spelling variants, acronyms, and abbreviations. Then, we developed I2E queries to translate documented mentions of stressors to structured data elements including the following: 1) subject ID, 2) Note ID, 3) note type, 4) provider type, and 5) mention of stress. Herein we report a descriptive summary of this NLP extraction process. The gold standard evaluation for NLP algorithm performance using a validation dataset is underway. We will report the evaluation results at the time of the conference.

Results: The I2E algorithms processed 197,346 clinical notes for 3,138 eligible patients within 18 seconds. Without negation, 2,452 (78.1%) patients have stress mention(s) detected from 23,367 clinical notes. With negation, 1,917 (61.1%) patients have stress mentions detected from 19,334 clinical notes. Among 145 terms of stressor, the most prevalent terms are depression/depressed, anxiety, stress, tired, and nervous. The most prevalent note types are progress notes, history and physical exam, consults, telephone encounter, and discharge summary. The most prevalent negations are stress test, stress incontinent, stress echo, stress dose, and ST depression.

Conclusions: Mental stressors are commonly documented in clinical narratives for patients with prostate cancer. Natural language processing can effectively extract mental stressors from EHR. This approach can potentially aid prospective data collection for mental stressors. However, a thorough evaluation of NLP algorithm performance is warranted as the next step.

Acknowledgement: This research was supported in part by NIH/NCATS grants (5UL1TR000062-05 and 1U54MD010706-01) to the Medical University of South Carolina.